PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
C12N 15/64, 15/66, C12Q 1/02, 1/68

A1

(11) International Publication Number: **WO 99/55886**

(43) International Publication Date: 4 November 1999 (04.11.99)

(21) International Application Number: PCT/US99/08823

(22) International Filing Date: 21 April 1999 (21.04.99)

(30) Priority Data:
09/065,775 24 April 1998 (24.04.98) US

(71) Applicant: GENOVA PHARMACEUTICALS CORPORATION [US/US]; 4233 Town Court North, Larenceville, NJ 08648 (US).

(72) Inventors: CEN, Hui; 5142 Masonic Avenue, Oakland, CA 94618 (US). SUN, Shaojian; 4233 Town Court North, Lawrenceville, NJ 08648 (US).

(74) Agents: ANTLER, Adriane, M. et al.; Pennie & Edmonds LLP, 1155 Avenue of the Americas, New York, NY 10036 (US).

(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: FUNCTION-BASED GENE DISCOVERY

(57) Abstract

The present invention relates generally to the field of genomics. More particularly, the present invention relates to methods for function-based gene discovery. Genes are identified as having or being associated with a specific function, as participating in a specific functional pathway, or as being a member of a specific functional group, by functional expression in one or more biological readout assays. This invention is based, at least in part, on the recognition that the signal-to-noise ratio of a readout assay used to screen a cDNA library can be significantly enhanced by methods which localize multiple molecular copies of each unique clone into discrete regions or compartments prior to functional expression. In one embodiment, this invention provides methods for *in situ* transfection of a sorted library in a "bar-coded" vector to carry out expression of genes from libraries being screened in readout cells. It is the ability to detect a biological readout in a readout cell line which enables the user to identify genes having specific functions. The methods set forth herein are suitable for application in a high throughput format for identification of genes and their functions simultaneously.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

FUNCTION-BASED GENE DISCOVERY

1. FIELD OF THE INVENTION

The present invention relates generally to the field of genomics. More particularly, the present invention relates to methods for function-based gene discovery. Genes are identified as having or being associated with a specific function, as participating in a specific functional pathway, or as being a member of a specific functional group, by functional expression in one or more biological readout assays. This invention is based, at least in part, on the recognition that the signal-to-noise ratio of a readout assay used to screen a cDNA library can be significantly enhanced by methods which localize multiple molecular copies of each unique clone into discrete regions or compartments prior to functional expression. In one embodiment, this invention provides methods for *in situ* transfection of a sorted library in a "bar-coded" vector to carry out expression of genes from libraries being screened in readout cells. It is the ability to detect a biological readout in a readout cell line which enables the user to identify genes having specific functions. The methods set forth herein are suitable for application in a high throughput format for identification of genes and their functions simultaneously.

2. BACKGROUND OF THE INVENTION

In the past 25 years, approximately 5,000 human genes have been cloned in full length and characterized by specific biological functions through various of assay systems. This represents only about 5% of an estimated 100,000 different genes in the human genome. A state-of-the-art method for determining gene function is to first individually clone a full-length cDNA encoding a protein-of-interest and then to perform assays in an attempt to determine biological function of the cloned gene. This approach can be quite expensive and inefficient due to the high cost of labor and materials for such work and because the entire process must be repeated from the beginning for each new gene to be characterized. In this regard, it generally takes several years for a skilled researcher to identify a new gene and characterize its corresponding function using methods focused on individual genes. With the advent of nucleic acid array technology to determine differential mRNA expression, the ability exists to analyze more than one gene at a time. For example, one might employ this method to select a subset of the 95,000 remaining genes to be characterized in a given tissue of interest. With this improvement alone, however, the speed

of gene function discovery would remain painfully slow. One reason is that differential mRNA expression analysis does not address the question of gene function in any depth. Instead, the technique may simply mark a gene as an interesting target worthy of further study. By contrast, functional expression in a heterologous background permits a gene to display a detectable function.

Some attempts have been made to develop systems for mammalian genetic functional screening (described further in Section 2.2 below). However, all of these attempts have involved detecting a positive signal through conferring a growth advantage. Such systems require weeks to grow cells under selective pressure, and the use of selection tends to result in the cloning of mutated genes. Of course, these systems are limited to functional identification of growth-related genes, and will not identify genes, for example, that are associated with cell death, that are toxic to a cell, or that cause other morphological changes. Accordingly, there is an urgent need for increased efficiency in the process of gene identification and functional characterization such that it not only takes less time but also yields more information.

2.1. GENOMIC SCIENCES AND DRUG DISCOVERY

There are about 40,000 prescription drug products currently available on the market, including over 6,700 brand names and 1,600 FDA-approved drugs. Despite this large number, drugs are known to work on only about 417 molecular targets in the human body and fewer than 80 molecular targets in bacteria, viruses and parasites (*see* Drews, 1996, Genomic sciences and the medicine of tomorrow, Nature Biotechnology 14, 1516-1518). Of course, drugs achieve their desired effects by binding to specific cellular targets (*e.g.* receptors, ion channels, enzymes, and other proteins or molecules). For example, breakthrough drugs for hypertension, depression, migraine, schizophrenia, and ulcers all act via specific receptors (Drews, 1996, *Id.*).

There are perhaps 100 to 150 major diseases in need of development of new treatments (Drews, 1996, *Id.*). If the number of genes contributing to each of these complex disease phenotypes is five to ten, and if each gene product interacts with from three to ten other gene products, then the number of genes associated with these conditions is perhaps from 3,000 to 10,000 (Drews, 1996, *Id.*). All disease-associated genes may be considered to be potential drug intervention targets and/or diagnostic markers. By contrast to the 420

known human molecular targets on which currently-available drugs are believed to work, the majority of the 3,000 to 10,000 human disease-associated genes have not yet been identified. From these considerations it is readily apparent that isolation and characterization of remaining disease-associated genes will dramatically broaden the horizon for development of new and/or improved treatments for most human diseases.

Identification and functional characterization of previously-unknown genes provides proteins which may be useful as drugs. Historically, such drugs have been useful when the body makes too little of an important protein, or when the presence of supplemental amounts of a protein can arrest or reverse a disease process. Protein drugs have been made possible by genetic engineering which has enabled industrial-scale protein production. Among the current beneficiaries of protein drugs are, *e.g.*, individuals who have had heart attacks and receive clot-dissolvers, individuals with renal failure who receive erythropoietin for anemia, and individuals with diabetes who receive recombinant human insulin.

Identification and functional characterization of new genes also provides reagents for potential use in gene therapy. Gene therapy may be defined as the introduction of genetic material into an individual for therapeutic benefit. Gene therapy may be used, for example, to correct detrimental genetic changes that occur in tumor cells, or to direct an individual's cells to produce a specific protein having therapeutic value. Although gene therapy still faces many technical hurdles, it offers promise for treatment of many disorders. Such disorders include those associated with single genes, such as hemophilia, sickle cell anemia, thalassemia, Gaucher's Disease, Huntington's chorea, and many others. More complex, polygenic diseases, such as diabetes and Alzheimer's disease, are also likely to benefit from gene therapy.

In more complex conditions, such as dementia and severe obesity, several distinct diseases may actually exist concurrently. In such cases, a condition may be mistakenly classified as a single disease simply because medical science lacks the information and tools necessary to distinguish among the underlying disease processes. If functional information were available for most genes in the genome, it might become possible to accurately identify each specific disease and a corresponding optimal therapeutic intervention within such complex conditions.

Discovery of new genes and their functions permits development of diagnostics for early detection of diseases. Such diagnostics, in turn, permit timely use of drugs or other

therapies for preventing irreversible damage. For example, current commercially-available gene-based diagnostics include tests for hemophilia A and B, phenylketonuria, retinoblastoma, and sickle-cell anemia. New, gene-based diagnostics may also be used to enhance the success rate of an existing therapeutic by identifying specific individuals within an affected group who respond well to a specific drug therapy. Similarly, diagnostics may help in development of new therapeutics through enhancement of understanding of differences among people in response to various medicines.

Existing expressed sequence tag (EST) databases are not, by themselves, sufficient to determine biological function. EST databases only suggest functional information to the extent that an EST encodes a domain of known function. Such databases do not provide any functional information for completely novel genes (*i.e.* genes not encoding any known domains or motifs).

As mentioned above, the state-of-the-art for determination of gene function has been to first clone a full length cDNA and then pursue functional characterization on an individual gene basis. The time consuming nature of the so-called single-gene approach can be illustrated by examination of the progress made. By 1995, the rate of functional characterization of newly-discovered genes reached a plateau of about 2,000 genes per year. If this rate continues, it would take another 46 years to identify the function of the genes remaining to be characterized in the human genome. The invention set forth herein provides methods to accelerate this schedule considerably.

It is believed the most efficient way to accomplish this characterization is to combine information from total genome sequencing with a database on gene expression patterns and another database on biological function, so that most of the estimated 100,000 genes encoded by the human genome can be grouped into a much smaller number of multi-component, core processes of known biochemical functions. Following this approach, gene groups, and then genes having strong medical relevance, would be prioritized for further, more thorough biological studies.

In contrast to the single-gene approach employed by previously available technologies, the invention described herein provides high throughput methods which combine the simultaneous isolation of gene structure with identification of gene function and/or functional gene group. By doing so, the method is able to directly screen mammalian cDNA libraries (average size 10^6 clones) using mammalian cell systems for

biological functions with specific cellular markers. With high resolution bioassay technology, this strategy also yields all genes in the human genome which are involved in a particular biological functional process of interest. In addition, this strategy makes it possible to automate the gene function screening process in a high throughput fashion. Accordingly, this invention provides a much more efficient way to characterize the function of the human genome, as described in detail in Section 5 below.

A brief overview of functional genomics approaches currently under way serves to illustrate the state of the art. With regard to EST databases, Human Genome Sciences, Inc. (HGS) and Incyte Pharmaceuticals, Inc. (Incyte), have produced proprietary EST databases comprising partial sequences of perhaps more than 70% of all human genes. Despite the fact that these EST databases are not yet linked in a meaningful way to functional information, seventeen of the largest pharmaceutical companies have spent more than $482 million to subscribe, according to a 1996 report (Friedrich, 1996, Nature Biotechnol. 14, 1234). Additional organizations have chosen positional cloning strategies for linking gene structure with function. Still other organizations are applying nucleic acid array technologies for analysis of expressed genes in a given tissue or cell (*e.g.* Affymetrix, Incyte).

Array technology, which represents the first attempt to go beyond single-gene methods of genome analysis, remains limited to characterization of gene expression as opposed to characterization of gene function. For example, one may use array technology to determine differential gene expression patterns in disease, thereby narrowing disease-gene candidates to a subset of genes. However, even under this approach, the speed of gene function discovery is not likely to increase significantly. This is so since such an approach would identify not only genes which may contribute to the cause of a disease process but also genes having altered expression as a consequence of a disease process. Further, the number of genes in the latter category is likely to vastly outnumber those in the first category. Analyzing potentially hundreds of genes that may be implicated in a given disease by an expression analysis using a single-gene approach would quickly become an overwhelming task. This is particularly evident when one considers that a given organization generally has a limited number of biological assays available in-house, *i.e.* far from enough for beginning to determine the biological function of new genes *en masse*.

It is clear that a genetic screen capable of identifying all genes associated with a specific biological function would be the most efficient way of linking gene structure to function. Although genetic screening approaches have been widely used for such organisms as *Drosophila* and *C. elegans*, there is no such approach widely applicable to mammalian systems. This is primarily due to the large size of mammalian genomes (*i.e.* 10^5 genes) and a lack of sensitive assay systems for detecting positive signals over background noise.

Nevertheless, limited attempts have been made at developing mammalian genetic functional screening systems. For example, such systems have been described by Deiss and Kimchi (1991, Science 252, 117-120) and by Cohen (1996, Cell 85, 319-329). However, these systems are slow, labor-intensive, restricted to cloning growth-related genes, and have a tendency to isolate mutated genes. This latter tendency arises from a requirement for relatively long-term culture (*i.e.* two or more weeks) under selective pressure for identification of a growth phenotype.

Accordingly, a great need exists for a large-scale (*i.e.* genome-wide) mammalian genetic functional screening method which may be employed over a time period of days instead of weeks and which provides an automated, general format for use instead of a manual, specific format that must be tailored to each functional readout assay. This invention provides such a method, as described in detail in Section 5 below.

2.2. EXPRESSION CLONING

Many methods have been described for cloning genes by functional expression. One method by Clarke *et al.* (June 23, 1987, Method for identification and isolation of DNA encoding a desired protein, U.S. Patent No. 4,675,285) provides a ten-step approach for selection of cDNAs expressed from sub-pools of a library which includes testing media from cultured cells in which sub-pools are expressed so as to identify a cDNA encoding a desired protein. Another method by King *et al.* (August 5, 1997, "Method of expression cloning," U.S. Pat. No. 5,654,150) provides an improvement which employs pools of about 100 individual bacterial colonies. Yet another method by Sang (March 31, 1993, "Expression cloning method," European Patent Application Pub. No. 0 534 619 A2) employs antibodies or ligands to screen expression libraries. As a general proposition, however, these methods have often been designed for very specific purposes, *i.e.* for identification of a single gene, and therefore lack general utility. For example, one method

utilized a transient COS cell expression assay and monoclonal antibody binding to identify CD28 (Aruffo and Seed, 1987, Proc. Natl. Acad. Sci. U.S.A. 84, 8573-8577).

3. SUMMARY OF THE INVENTION

This invention provides methods for function-based gene discovery. Genes are identified as having or being associated with a specific function, as participating in a specific functional pathway, or as being a member of a specific functional group, by functional expression in one or more biological readout assays. This invention is based, at least in part, on the recognition that the signal-to-noise ratio of a readout assay used to screen a cDNA library can be significantly enhanced by methods which localize multiple molecular copies of each unique clone into discrete regions or compartments prior to heterologous expression. In one embodiment, this invention provides methods for *in situ* transfection of a sorted library in a "bar-coded" vector to carry out expression of genes from libraries being screened in heterologous readout cells. It is the ability to detect a biological readout in heterologous cells which enables the user to identify genes having specific functions. The methods set forth herein are suitable for application in a high throughput format for identification of genes and their functions simultaneously.

This invention provides a method for enhancing the signal-to-noise ratio of a biological readout assay used to screen a bar-coded cDNA library comprising: (a) sorting the bar-coded cDNA library using a nucleic acid array; and (b) transfecting the library sorted in step (a) into a readout cell line *in situ*. In one embodiment, the nucleic acid array is a biological array or a gene chip. in another embodiment, the biological array comprises a vector carrying a plurality of complementary bar codes. In still another embodiment, the plurality of complementary bar codes is immobilized on a support. In a preferred embodiment, the support is nitrocellulose or nylon. In another embodiment, transfecting *in situ* is carried out using a chemical transfectant or electroporation. In another embodiment, the readout cell line is NIH 3T3 cells carrying a reporter gene under the control of a response element or promoter. Selection of the response element or promoter is guided by the particular readout assay selected. In still another embodiment, the reporter gene is selected from the group consisting of β-galactosidase, luciferase and chloramphenicol acetyltransferase. In a preferred embodiment, the response element or promoter is selected from the group consisting of an NFκB response element, an NFAT response element, a

cyclic adenosine monophosphate response element, a STAT-inducible promoter, a LEF-1-inducible promoter and a p53-inducible promoter. In another preferred embodiment, the cDNA library is tetracycline-inducible or estrogen inducible. In still another preferred embodiment, the biological readout assay detects genes in a pathway selected from the group consisting of a mitogenic signaling pathway, a STAT signaling pathway, an NFκB signaling pathway, a stress signaling pathway, an apoptosis signaling pathway, an NFAT signaling pathway, a Wnt signaling pathway, a CREB signaling pathway, an AP-1 signaling pathway, a proliferation signaling pathway and an anti-proliferation signaling pathway.

This invention provides a method for conducting a biological readout assay used to screen a bar-coded cDNA library comprising: (a) sorting the bar-coded cDNA library using a nucleic acid array; (b) transfecting the library sorted in step (a) into a readout cell line *in situ*; and (c) conducting the biological readout assay. In another embodiment, the nucleic acid array is a biological array or a gene chip. In another embodiment, the biological array comprises a vector carrying a plurality of complementary bar codes immobilized on a support. In another embodiment, the plurality of complementary bar codes consists of from 10^2 to 10^8 complementary bar codes. In another embodiment, the support is nitrocellulose or nylon. In another embodiment, transfecting *in situ* is carried out using a chemical transfectant or electroporation. In another embodiment, the readout cell line is NIH 3T3 cells carrying a reporter gene under the control of a response element or promoter. In another embodiment, the reporter gene is selected from the group consisting of β-galactosidase, luciferase and chloramphenicol acetyltransferase. In another embodiment, the response element or promoter is selected from the group consisting of an NFκB response element, an NFAT response element, a cyclic adenosine monophosphate response element, a STAT-inducible promoter, a LEF-1-inducible promoter and a p53-inducible promoter. In another embodiment, the bar-coded cDNA library is tetracycline inducible or estrogen inducible. In another embodiment, the biological readout assay is capable of detecting genes in a pathway selected from the group consisting of a mitogenic signaling pathway, a STAT signaling pathway, an NFκB signaling pathway, a stress signaling pathway, an apoptosis signaling pathway, an NFAT signaling pathway, a Wnt signaling pathway, a CREB signaling pathway, an AP-1 signaling pathway, a proliferation signaling pathway and an anti-proliferation signaling pathway.

This invention provides a method for conducting a biological readout assay used to screen a bar-coded cDNA library comprising: (a) sorting the bar-coded cDNA library using a nucleic acid array having a plurality of concave loci; (b) expressing the bar-coded cDNA library sorted in step (a) using *in vitro* transcription and translation to produce a population of proteins; and (c) screening the population of proteins produced in step (b) for a biochemical activity-of-interest, so as to conduct the biological readout assay. In one embodiment, the biochemical activity-of-interest screened in step (c) is selected from the group consisting of a receptor-binding activity, a ligand-binding activity and a growth factor activity. In another embodiment, screening is carried out by immobilizing the population of proteins on a solid support for use in a binding assay. In another embodiment, the solid support is nitrocellulose or nylon. In another embodiment, screening is carried out by placing the population of proteins in contact with readout cells for use in a biological activity assay.

This invention provides a method for identifying one or more genes-of-interest in a pre-sorted cDNA library comprising: (a) transfecting the pre-sorted cDNA library into a population of readout cells; and (b) screening the population of readout cells transfected in a biological readout assay, to identify one or more genes-of-interest.

In one embodiment, the pre-sorted cDNA library comprises a bar-coded cDNA library hybridized to a nucleic acid array. In another embodiment, transfecting is carried out using chemical transfectants or electroporation. In another embodiment, the biological readout assay identifies one or more genes-of-interest in a pathway selected from the group consisting of a mitogenic signaling pathway, a STAT signaling pathway, an NFκB signaling pathway, a stress signaling pathway, an apoptosis signaling pathway, an NFAT signaling pathway, a Wnt signaling pathway, a CREB signaling pathway, an AP-1 signaling pathway, a proliferation signaling pathway and an anti-proliferation signaling pathway.

This invention provides a method of expression cloning one or more genes-of-interest in a cDNA library comprising: (a) sorting the cDNA library; (b) transfecting the sorted library into a readout cell line; and (c) identifying a positive signal from the transfected library in a biological readout assay, so as to expression clone one or more genes-of-interest in the cDNA library. In one embodiment, sorting the cDNA library is carried out using a nucleic acid array. In another embodiment, transfecting the sorted library is carried out using chemical transfectants or electroporation.

the positive signal is identified by immunocytochemistry. In another embodiment, the biological readout assay identifies one or more genes-of-interest in a pathway selected from the group consisting of a mitogenic signaling pathway, a STAT signaling pathway, an NFκB signaling pathway, a stress signaling pathway, an apoptosis signaling pathway, an NFAT signaling pathway, a Wnt signaling pathway, a CREB signaling pathway, an AP-1 signaling pathway, a proliferation signaling pathway and an anti-proliferation signaling pathway.

This invention provides a method of sorting a cDNA library for use in an expression cloning assay comprising: (a) cloning a population of cDNA inserts into a population of bar-coded vectors; (b) preparing the population of bar-coded vectors for hybridization to a DNA array by exposing only the bar code region in single-stranded form; and (c) hybridizing the population of bar-coded vectors to a nucleic acid array to sort the cDNA library. In one embodiment, the nucleic acid array is selected from the group consisting of a gene chip and a biological array. In another embodiment, preparing the population of bar-coded vectors for hybridization to a DNA array by exposing only the bar code region in single-stranded form in step (b) is carried out using the following steps in the order stated: (a) digesting the population with a restriction endonuclease to linearize the population; (b) binding a DNA-binding protein to at least two sites on the population; and (c) digesting the population bound in step (b) to expose the single-stranded bar code region. In another embodiment, the DNA-binding protein is selected from the group consisting of a lactose repressor protein, a tetracycline repressor protein, E2F, AP1, SP1 and p53. In another embodiment, the restriction endonuclease is selected from the group consisting of NotI, SfiI and EcoRI. In another embodiment, digesting the vector population in step (c) is carried out using an enzyme selected from the group consisting of exonuclease III, T4 DNA polymerase, Klenow fragment, T7 DNA polymerase, Vent DNA polymerase and Pfu DNA polymerase.

4. BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1. A phagemid vector for making a bar-coded cDNA library.

FIG. 2A-2B. 2A. Preparation of a bar-coded vector. 2B. Preparation of a bar-coded cDNA library.

FIG. 3. Sorting a bar-coded cDNA library.

FIG. 4. Flow chart of gene identification methods from the step of *in situ* transfection to the step of cDNA retrieval.

FIG. 5. Illustration of a gene chip with a plurality of concave loci.

5. DETAILED DESCRIPTION OF THE INVENTION

This invention provides methods for function-based gene discovery. Genes are identified as having or being associated with a specific function, as participating in a specific functional pathway, or as being a member of a specific functional group, by expression in one or more biological readout assays. This invention provides expression cloning methods enabling high-throughput library screening for determination of gene function. The invention is based, at least in part, on the recognition that the signal-to-noise ratio of a readout assay used to screen a cDNA expression library can be significantly enhanced by localizing multiple molecular copies of each unique clone into discrete regions or compartments. It is the ability to detect a biological readout in heterologous cells which enables the user to identify genes having specific functions. A major advantage of the invention is to provide methods for assaying all genes in a cDNA expression library simultaneously, instead of one-at-a-time, under conditions in which the readout signal-to-noise ratio is significantly enhanced. Moreover, a rational basis for characterization of functional gene groups is provided where more than one gene is identified in any given readout assay.

In one embodiment, this invention provides methods for *in situ* transfection of a sorted library in a "bar-coded" vector to carry out expression of genes from libraries being screened in heterologous readout cells. The vector "bar code" is an oligonucleotide

sequence within the vector which is unique to each individual clone of a library. The bar code enables sorting of the library in physical space by hybridization to nucleic acid arrays which are complementary to library bar code sequences. The bar code unique to each clone, together with the unique position of each complementary bar code in a nucleic acid array, provides a method for direct retrieval of a gene having a function of interest identified in any given readout assay. Moreover, each unique bar code can serve as a specific primer for PCR and/or sequencing of a desired clone in a library.

5.1. GENERAL CONSIDERATIONS

In both above-mentioned embodiments of the invention, it is the ability to detect a biological readout upon heterologous expression which enables the user to identify genes having specific functions. These embodiments are described in detail in Sections 5.2 below. A major advantage of the invention is to provide methods for assaying all genes in a cDNA library simultaneously for the ability to modify a specific biological function associated with a specific readout assay. The pattern of gene activity in any given readout assay also provides a rational method for identification of functional gene groups. The methods set forth are suitable for application in a high throughput format for rapid identification of genes and their functions. For example, such a high throughput format may easily screen, at least 10^4, or from 10^4 to 10^6, independent recombinants for functional activity at one time.

To practice the invention, a complementary DNA (cDNA) library is prepared from messenger RNA (mRNA) obtained from a cell population of interest (*e.g.* a cell population may be derived from a specific tissue, disease, or biological state). A cDNA library may also be purchased commercially. The cDNA is operably linked to an expression vector suitable for use with the invention. Constructs are prepared and purified using standard recombinant DNA techniques as described in, *e.g.*, Sambrook *et al.* (1989, Molecular Cloning: A Laboratory Manual, 2nd Ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York). Expression vectors suitable for use with the invention are available commercially, or can be specially designed by the user or as described herein.

5.1.1. OVERVIEW

The library technology of the invention has, *inter alia*, the following three features: (a) inducibility of library gene expression; (b) suitability for use with sense and antisense libraries; and (c) suitability for use with libraries from various disease and non-disease tissues and/or cells, and/or various stages of development of interest. The user will note that when a bar-coded vector is used virtually any cDNA library vector is made suitable when modified to comprise a bar code as described herein. Such vectors do not require an inducible promoter because these cDNA libraries are directly transfected into readout cells without having to be propagated in virus-producing cells.

The method is suitable for use with a microscope-based, *in situ* approach for detection of various readouts. Such readouts may include, but are not limited to: target protein expression; target mRNA expression; cellular localization changes; and/or cellular morphology changes. Such single-format, microscope-based detection may be easily automated. Because screening of libraries and/or sub-libraries requires only a few days, the possibility of appearance of mutated genes during prolonged growth in cell culture, as with prior art methods, is largely eliminated.

The methods of the invention are suitable for use in high throughput screening of a large number of functional (*i.e.* readout) assays of interest. Such functional assays include cell culture-based assays (*i.e.* cellular readout assays, see Section 5.5) that rely upon expression of genes in a library and detection of a functional effect of expression. This accelerated time scale provides a two-fold advantage in that it (a) requires a reduced workload relative to procedures requiring longer assay time; and (b) vastly reduces the appearance of mutated genes arising from prolonged cell culture time. The functional assay technology that can be used includes all existing immunostaining assays and biochemical assays. See Sections 5.5-5.7 below for a description of assays and Section 6 below for examples. Such assays are designed to identify genes involved in major disease categories as well as genes that regulate various cellular physiological functions.

5.1.2. mRNA SOURCES

There are no special considerations when choosing a messenger RNA (mRNA) source for construction of a cDNA library for use with the methods of the invention. Any mRNA source may be used. Accordingly, cells suitable as sources of mRNA from which a cDNA library may be constructed include, but are not limited to, mammalian cells, bacterial cells, yeast cells, insect cells and amphibian cells. However, because of (a) the relative absence of genetic functional screening systems available for mammalian organisms compared to, say, flies or yeast, and (b) the relative complexity of the mammalian genome, the methods of the invention are preferred for use in screening mammalian cDNA libraries. Suitable mammalian mRNA sources include tissues and cell lines. Mammalian tissues that may be used include normal and disease tissues (*e.g.* carcinomas, lymphomas). Mammalian cell lines that may be used include any of the cell lines available from the American Type Culture Collection (ATCC). Exemplary mammalian cell lines include Chinese hamster ovary (CHO) cells, HeLa cells, baby hamster kidney (BHK) cells, monkey kidney cells (*e.g.* COS), human hepatocellular carcinoma cells (*e.g.* Hep G2), human embryonic kidney cells (*e.g.* HEK 293), mouse sertoli cells, canine kidney cells (*e.g.* MDCK), buffalo rat liver cells, human lung cells, human liver cells and mouse mammary tumor cells.

5.1.3. cDNA LIBRARIES

Sense or antisense cDNA libraries may be generated by any method known in the art. Many such methods exist and examples may be found in Sambrook et al. and Ausubel et al., both of which are incorporated by reference herein in their entireties (Sambrook et al., 1989, Molecular Cloning, A Laboratory Manual, 2d Ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York; Ausubel et al., eds., in the Current Protocols in Molecular Biology series of laboratory technique manuals, © 1987-1997 Current Protocols, © 1994-1997 John Wiley and Sons, Inc.). Many references are available which describe antisense cDNA library construction (*see e.g.*, Spann et al., 1996, Proc. Natl. Acad. Sci. U.S.A. 93, 5003-5007; and Deiss and Kimchi, 1991, Science 252, 117-120).

The library may be an antisense library such that antisense polynucleotides are generated upon expression of the library. Such antisense polynucleotides may, for example, provide a source of inhibition of a detectable cellular event in a functional assay. In this

way, an antisense expression library will identify one or more genes required for operation of a specific pathway by "knocking out" (*i.e.* rendering inoperative) such a pathway.

A library may be divided into subpools for screening. For example, from 100 to 1,000 subpools may be generated, each subpool comprising a cDNA diversity of from about 10,000 to about 1,000 clones, thus representing a cDNA diversity in all pools combined of about 1,000,000. Each library subpool may be individually (*i.e.* separately) expressed in a heterologous cell population.

A library may be a normalized cDNA library. Any cDNA library normalization technique known to one skilled in the art may be used with the methods of the invention. For example, see "Normalization and subtraction: two approaches to facilitate gene discovery," Genome Research 6, 791-806 (1996).

5.2. BAR-CODED VECTORS

A "genetic bar code" is an oligonucleotide tag or label having a specific sequence. This invention provides a method of constructing a cDNA library in a vector containing a plurality of genetic bar codes at a diversity equal to or larger than the diversity of the cDNA library. This invention provides methods for sorting and transfecting such a library. The methods employ a unique genetic bar code linked to each clone in a library for various uses (*e.g.* sorting, retrieval of insert).

The human genome is believed to encode about 100,000 genes, and any given human cell or tissue may express from about 10,000 to about 50,000 of these genes. Therefore, in order to cover every expressed gene (including rare genes) during preparation of a human cDNA library from messenger RNA, it is preferred that about 10^6 independent clones be used. Accordingly, a vector having about 10^6 unique genetic bar codes is preferred since it is preferred that a unique genetic bar code be associated with each library clone.

A bar code having ten nucleotides and using all four possible bases at each position is capable of generating a set of genetic bar codes having a diversity of 4^{10} or 1.048×10^6. The optimum length and base composition of oligonucleotides (oligos) for specific and efficient hybridization to complementary sequences may be chosen by the user. For example, oligos 15 to 20 nucleotides long having a diversity of 4^{15} to 4^{20} (*i.e.*, 10^9 to 10^{12}) may be used to cover a library of 10^6 diversity to ensure that any two genetic bar codes are

different by several nucleotides. Various approaches known in the art may also be used to reduce cross-hybridization among different bar codes (*see e.g.* Shoemaker et al., 1996, Nature Genetics 14, 450-456).

5.2.1. CONSTRUCTING A BAR-CODED VECTOR

The vector employed for use with a bar-coded cDNA library may be any vector incorporating a genetic bar code. In one embodiment, a suitable vector comprises the genetic bar code and a eukaryotic promoter. In another embodiment, a suitable vector comprises the genetic bar code, a eukaryotic promoter (*e.g.* a CMV promoter), a cDNA insert, an f1 origin, an antibiotic resistance gene (*e.g.* ampicillin resistance gene), an SV40 origin and a ColE origin (*see e.g.* FIG. 1). For the phagemid vector illustrated in FIG. 1, sites 1 and 2 may be used for inserting the genetic bar code, sites 3 and 4 may be used for inserting cDNA, the f1 origin may be used for making single-stranded DNA, and the antibiotic resistance gene provides for growth selection of the phagemid vector in *E coli.* The ColE and SV40 origins may be used to provide high copy number amplification of phagemid DNA in bacteria and eukaryotic cells, respectively (*see* FIG. 1).

By way of example and not limitation, a bar-coded library vector may be constructed as illustrated in FIG. 2A. Here, the vector and the bar code mixture are each digested with enzymes 1 and 2 (which cut at sites 1 and 2, respectively), and then ligated to produce the bar-coded library vector (FIG. 2A). A bar-coded library may be constructed as illustrated in FIG. 2B. Messenger RNA (mRNA) is reverse transcribed using an oligo-dT primer containing restriction site 3 using methods well known to those skilled in the art (*see* FIG. 2B). Following conversion into double-stranded cDNA, an adapter or linker containing restriction site 4 is ligated to the 5' end (relative to the sense strand). The resulting double-stranded cDNA bears site 4 at its 5' end and site 3 at its 3' end. This cDNA and the bar-coded vector are each digested with enzymes 3 and 4 and ligated together to produce the double-stranded, bar-coded cDNA library. It is preferred that any library amplification be performed on plates, as opposed to in solution, to ensure equal amplification of all clones represented.

One skilled in the art will readily recognize and appreciate that the various features of a suitable vector need not be precisely as illustrated. For example, the location of the bar code sequence can be other than at the illustrated location within the vector.

To ensure that each bar-coded vector only carries one genetic bar code, a population of double-stranded genetic bar codes is designed in dephosphorylated form and having a staggered restriction enzyme site at both ends of the bar code (*e.g.* EcoRI). For example, where chemically-synthesized oligonucleotides are already dephosphorylated following the chemical synthesis. An enzyme (*e.g.* alkaline phosphatase) can also be used to insure the dephosphorylated state. This method ensures that, after annealing of the two bar code strands, a double-stranded bar code is formed which lacks the phosphorylation which would be necessary for the formation of a bar code dimer.

Further, one can apply a "zero background" cloning system (*e.g.*, such a system is commercially available from InVitrogen) to clone a bar code population in to the chosen vector. A zero background cloning system is a positive selection system for prokaryotic cloning which works by direct selection of inserts via disruption of the lethal gene *ccdB* (control of cell death). In such a system, only bacteria transformed with a genetic bar code inserted into the vector will survive and be propagated. In this way, the vector population generated will not contain any individual vectors lacking a bar code.

5.2.2. SORTING A BAR-CODED LIBRARY

Sorting of a bar-coded library may be carried out using supports having bound thereto oligonucleotide sequences that are complementary to the genetic bar codes of the cDNA library. A DNA sequence complementary to each genetic bar code in a bar-coded library is affixed (*e.g.* deposited or synthesized) at discrete locations of a nucleic acid array. Natural or modified nucleotides can be used for synthesis. Use of certain modified nucleotides may promote formation of stronger bonds with the complementary bar code of a vector. In this regard, the bonding properties of common modified nucleotides such as phosphorothioates have been well described.

An array may be a commercially available gene chip (*e.g.* Affymatrix, Incyte) or may be manufactured using methods known in the art. Many such methods have been described (for a brief review, *see* Ramsay, January 1998, Nature Biotechnol. 16, 40-44). For example, light-directed, solid-phase synthesis technology permits massive numbers of oligonucleotides to be synthesized on a support at precise positions (*see e.g.* Fodor *et al.*, August 29, 1995, U.S. Pat. No. 5,445,934). To achieve gene separation and sorting using such an array, the array is hybridized with the single stranded genetic bar code region of a

double stranded vector (*see e.g.* "ssGBC cDNA library" in FIG. 3, which illustrates such a single stranded genetic bar code region on a double stranded vector). In FIG. 3, a gene chip having a plurality of complementary genetic bar codes is shown. Each area of the chip (labeled A, B, C, and N) contains multiple copies of each unique complementary bar code. Following hybridization, each unique cDNA corresponding to each unique bar code is sorted into a discrete area on the chip (*see* bottom panel of FIG 7.)

Optimum hybridization conditions are used to ensure accurate base pairing between the various genetic bar codes of a library and their corresponding complements in a nucleic acid array so as to prevent or minimize mismatches. The hybridization (a) separates library vector molecules encoding distinct recombinants from each other and (b) sorts all library vector molecules encoding the same recombinant to discrete, known locations. This separation and sorting operation results in an equal amount of DNA at each location. Abundant genes of a library will be represented at multiple locations on a chip while rare genes will be represented at only one or a few locations. Equivalent amounts of DNA are hybridized at each location.

As an alternative to gene chip arrays, which may be expensive to manufacture, a "biological array" may be manufactured and used to sort a bar-coded expression library. Such a biological array is created from a library of sequences which are complementary to the bar codes of the expression vector. The biological array is segregated into discrete locations using the physiology of the microbe (*e.g.* bacteria or yeast), as follows. Since each microbe which takes up a plasmid DNA containing a complementary bar code will only retain a single type of plasmid, each complementary genetic bar code is automatically separated from all others at this step. The user will note that the vector chosen to produce the complementary bar code array is different from the expression vector used to create the library so as to preclude any hybridization between the two vectors.

In this way, microbial colonies carrying a plasmid library encoding complementary bar codes are used to construct a biological array. Such a biological array can be easily reproduced by replica plating. The DNA of the array is easily immobilized on a solid support (*e.g.* nitrocellulose, nylon, etc.) by well known methods. The sequence of the complementary genetic bar code at each location of the array may be determined by standard sequencing reactions. This information may then be stored in a computer for later retrieval as needed.

A biological array is different from that of a gene chip array as follows. Instead of having each complementary bar code present as a homogeneous nucleic acid at each location of the array, each is present in a background of other DNA (*i.e.* from the plasmid carrying the complementary bar code and the microbial genome).

5.2.3. *IN SITU* TRANSFECTION OF A SORTED LIBRARY

Analysis of a sorted library by functional expression provides a means to screen a large number of genes simultaneously in order to identify genes having the biological function specified by the chosen readout assay. Any of the following methods may be used for *in situ* transfection of a sorted library. Following all of the transfection procedures described below, readout cells are rinsed in a physiologic buffer and cultured for a period of time to allow expression of the transfected genes. The period of time to allow expression may be from 12 hours to 12 days, from 1 day to 6 days, or from 2 days to 3 days. In a preferred embodiment, the period of time is from 1 day to 4 days.

In one embodiment, *in situ* transfection may be performed using a gene chip having a plurality of concave loci (*i.e.* U-shaped areas), each locus having an oligonucleotide complementary to a genetic bar code attached thereto. Following library sorting by hybridization, such a gene chip will have an individual cDNA recombinant at each locus. *In situ* transfection is performed by contacting the chip to a readout cell culture in the presence of a solution which facilitates the release of the hybridized recombinants. Such a solution may be, *e.g.*, phosphate-buffered saline or tissue culture medium without serum supplement. Generally, any low-salt solution (*e.g.* 150 mM NaCl or lower) will result in the dissociation (*i.e.* release) of the hybridized recombinants from the gene chip. Chemical transfectants may also be included in the solution to facilitate uptake of the released DNA into the readout cells. Such transfectants may be any transfectant known in the art. For example, calcium phosphate, DEAE-dextran, polybrene or a lipid-based transfectant such as LT1 (Panvera) or Lipofectamine (GibcoBRL) may be used.

In another embodiment, *in situ* transfection may be performed using a biological array or a gene chip having a flat surface. Here, *in situ* transfection may be performed by contacting the biological array or the gene chip to a readout cell line in the presence of a solution as described above. A chemical transfectant may also be used as described above. In a preferred embodiment, a micro-compartmentalization grid device may also be used to

restrict diffusion of each released recombinant. Such a micro-compartmentalization grid device may be as illustrated in FIG. 3 of copending U.S. Patent Application No. 09/065,776, filed April 24, 1998, entitled "MICRO-COMPARTMENTALIZATION DEVICE AND USES THEREOF," by Cen and Sun (Attorney Docket No. 9557-003), which is incorporated herein by reference in its entirety.

In yet another embodiment, *in situ* transfection may be performed by electroporation. Electroporation may be performed using, *e.g.*, a cell culture device as described in U.S. Patent No. 5,134,070, which is incorporated by reference herein in its entirety. Here, the readout cell line used for electroporation may be any readout cell line which will attach to and proliferate on a solid support. Such a cell line may be grown in a monolayer on the bottom of a cell culture device which is electrically conductive (*see e.g.* U.S. Patent No. 5,134,070). A gene chip or biological array having a sorted cDNA library attached thereto is contacted with the cell line in the presence of a suitable electroporation solution (*e.g.* phosphate buffered saline or as described in U.S. Patent No. 5,134,070) such that it is between the electrically conductive upper and lower surfaces of the culture device. The contact of the upper surface of the culture device with the electroporation solution provides a continuous electric circuit for passing a current which mobilizes the DNA from the gene chip or biological array to the readout cells.

In yet still another embodiment, *in situ* transfection may be performed using enzymes to facilitate attachment to and release from a gene chip or a biological array. Here, one may covalently attach a sorted library to a nucleic acid array to ensure tight binding using T4 ligase, or an enzyme having a similar activity, to ligate the oligonucleotides encoding the complementary bar codes to the bar-coded cDNA library following hybridization. Such a covalently-bound bar-coded cDNA library may be released at will by including a restriction endonuclease in the transfection solution (*e.g.* if an EcoR1 site is used as site 2, *see* FIG. 1, then one may cut with EcoR1).

Finally, to ensure specific hybridization of a bar-coded cDNA library to a nucleic acid array, enzymes which cut mismatched nucleotides (such as T4 Endonuclease VII, also called resolvase, *see* Youil *et al.*, 1996, Genomics 32:431) may be used to eliminate cross-hybridization following the completion of the hybridization process.

An overview of the *in situ* transfection (gene transfer) methods as they relate to the overall process of gene identification and retrieval, is illustrated schematically in FIG. 4.

5.2.4. BIOCHEMICAL ANALYSIS OF A SORTED LIBRARY

In addition to expression of a sorted cDNA library in one or more cellular readout assays, biochemical analysis of a sorted library may also be performed. A solid support having a plurality of concave loci may be used, as described above and depicted in FIG. 5. For biochemical analysis, each individual protein encoded by the sorted library is first expressed using the well-known techniques of *in vitro* transcription and translation. Since each individual protein expressed is compartmentalized at a discrete locus, each may be screened for any biochemical activity of interest (*e.g.* receptor-binding activity, ligand-binding activity, growth factor activity). See Section 5.6 below for a description of various biochemical readout assays.

Individual proteins may be subsequently immobilized on a solid support (*e.g.* nitrocellulose or nylon) for use in binding or other assays. Alternatively, individual proteins may be left free within U-shaped wells for subsequent assay of activity. For example, for detection of growth factor activity, in vitro translation products may be placed in contact with readout cells using mild centrifugation to transfer the contents of each U-shaped locus onto a readout cell grid.

5.2.5. GENE RETRIEVAL AND MONITORING

Multiple methods are available for retrieval and monitoring of library inserts using genetic bar codes. Following identification of a specific clone-of-interest in a readout assay, the unique genetic bar code situated in the vector next to the cDNA insert may be used to isolate the clone-of-interest. For example, localized releasing of DNA hybridized on a nucleic acid array may be carried out by competition using an oligonucleotide identical to the bar code of interest. Further, isolation of the clone-of-interest may be carried out by polymerase chain reaction (PCR) to amplify only insert cDNA linked to the identified genetic bar code. Under this approach, for example, the bar code sequence may be used as a specific primer together with a suitable vector primer and total library DNA as template. The primer of the genetic bar code can also be used for isolating the specific plasmid by a procedure referred to as "gene trapping" (Gibco BRL). Briefly, "gene trapping" is a method for rapid isolation of cDNA clones from single stranded DNA prepared from a library. This method is based on isolating cDNA clones which hybridize with a biotinylated oligonucleotide complementary to a cDNA of interest (*see* Le *et al.*, 1995, Focus 17, 45).

Still further, isolation of the clone-of-interest may be carried out by physical "picking" since the location of each unique bar code sequence within an array is generally known.

The genetic bar code technology not only enables performing cellular functional assays on all genes represented in a cDNA library simultaneously, but is also amenable to automation. Such automation allows many functional assays to be performed in a single format, as described further in Section 5.8 below.

5.2.6. GENETIC BAR CODE DESIGN

To facilitate uniform hybridization between a bar coded library and its complementary sequences of a nucleic acid array, it is preferred that all oligonucleotides (*i.e.* genetic bar codes) have the same or nearly the same melting temperature (Tm). It is also preferred that conditions be provided in which only the genetic bar code of a bar-coded vector is in single stranded form, while the remainder of the vector remains in double stranded form, so as to minimize interactions among vectors carrying different but possibly related cDNA inserts. A bar code in a double-stranded vector can be exposed in single-stranded form using an enzyme having 3' to 5' exonuclease activity, such as T4 DNA polymerase, and a bar code population having one nucleotide omitted from all bar code sequences. Such an exonuclease activity is capable of cleaving nucleotides from a 3' recessed end, and such enzymes will cease exonuclease activity under certain conditions. For example, if C is omitted from the bar code sequence, then T4 polymerase may be used in the presence of G and the linearized double stranded vector to expose the bar code in single stranded form. As a further example, if A is omitted from the bar code sequence, T4 polymerase may be used in the presence of T and the linearized double stranded vector to expose the bar code in single stranded form. In other words, during a T4 exonuclease digestion, all nucleotide triphosphates (NTPs) are omitted from solution except the NTP complementary to the nucleotide omitted from the bar code. In this way, T4 DNA polymerase and a bar code population lacking one of the four nucleotides in its sequence may be used to make the bar code single-stranded and protruding from the end of a linearized, double-stranded vector. Alternatively, such an enzymatic exonuclease may be used when all four nucleotides are present in the bar code so long as the timing of the reaction is closely controlled to expose only the bar code in single stranded form.

In one embodiment, a genetic bar code pool may be designed using a set of nucleotide dimers as building blocks for synthesizing the pool. For example, one nucleotide dimer set consists of TG, AG, GA and GT. It is notable that this set has a minimum of one nucleotide identity between any two given dimers. Such a set is referred to herein as a minimal mismatch set (MMS). The average pair-wise mismatch for the set of TG, AG, GA and GT is 1.7 nucleotides, or 83%. The average pair-wise mismatch is computed by adding all the possible pair-wise nucleotide differences and dividing by the total number of pairs. In this case, it is 10/6, or about 1.7 nucleotides, or 1.7/2 = 85%. While the omitted nucleotide in the MMS listed above is C, the omitted nucleotide may be any of the four nucleotides when designing such an MMS. Nucleotide dimers are chosen such that the Tm for each dimer remains constant. In this way, the Tm for each genetic bar code of a pool will be the same. Methods for computing Tm are well known to one skilled in the art. The omission of a nucleotide in design of a bar code pool may be used to allow formation of a protruding end encoding the bar codes, as described above. A pool of 20-mers generated through random synthesis using the above-listed set of nucleotide dimers will produce a pool having a diversity of 4^{10} or 1.05×10^6 genetic bar codes. The minimum percentage of pair-wise mismatches within this pool of genetic bar code 20-mers is 1/20 or 5%, while, as noted above, the average pair-wise mismatch between any two genetic bar codes is 83%.

In another embodiment, genetic bar codes may be designed using a set of nucleotide trimers, each trimer having a minimum of two nucleotides different from any other trimer and each trimer having one G. Such an example set, which omits C, consists of AGT, TGA, TAG, ATG, GTA and GAT. The average pair-wise mismatch between any two genetic bar codes produced using this MMS is 2.4 nucleotides, or 80%. An oligonucleotide pool of genetic bar codes constructed randomly from eight rounds of synthesis using the above-listed six trimers will have a diversity of 1.68×10^6 with each bar code having a length of 24 nucleotides. The minimum percentage of pair-wise mismatch within this pool of genetic bar codes is 2/24 or 8.3%, while the average pair-wise mismatch between any two genetic bar codes in this set is 80%.

If 4-mer oligonucleotides are used as building blocks, each having a minimum of three nucleotides different at each position from all other 4-mers and each having one G, then there are eight building blocks in the MMS, as follows: GATT, TGAT, TAGA, TTTG, GTAA, AGTA, ATGT and AAAG (*see also* U.S. Patent No. 5,635,400 by Brenner).

Alternatively, one can choose another MMS from a total of 32 possibilities (*i.e.* $2^3C^1_4$ where C designates combinatorial) of 4-mers having one G nucleotide and 3 other nucleotides (T, A, or mixture of T and A). For example, an MMS may consist of GATA, GTAT, AGAA, TGTT, AAGT, TTGA, ATTG and TAAG. A pool of genetic bar codes constructed using seven 4-mer subunits will produce a bar code length of 28 nucleotides and a bar code diversity of 1×10^6. The minimum percentage of pair-wise mismatches within this genetic bar code pool is 3/28 or 10.7%.

In general, if an N-mer oligonucleotide building block having a number of G nucleotides in the N-mer equal to k and the remaining nucleotides (*i.e.*, N-k) consisting of A or T, or an A plus T mixture, then the bar code diversity is equal to $2^{(N-k)}C^k_N$. Further, there exists $2^{(N-k)}C^k_N$ number of MMS in these total possible constructs for a give minimal mismatch cut-off. In general, the number of sequences in different MMS is not the same. For example, in the case 5-mer with two G in each sequence and three nucleotides of A or T or A and T mixture. There are 80 sets of MMS. Some of these MMS have 8 sequences and some have 12 sequences for a mismatch cut-off of 3. It is generally true also that the larger the N for a given minimal mismatch cut-off, the larger the number of sequences in any set of MMS. It is also true that for a given diversity number in a library of genetic bar codes constructed from N-mer nucleotide subunits (such as those listed above, 8 tetrameric MMS in 4-mer example), the larger the N, the higher the minimal mismatch percentage in the library.

Accordingly, another way of constructing a pool of genetic bar codes is to use a certain number (e.g. 100) of oligonucleotides as building blocks selected from all possible combinations of a fixed length (*e.g.* 9-mer) with a certain minimum number (*e.g.* four) of nucleotides different among them. The diversity of genetic bar codes will be precisely 1,000,000 if the bar codes are composed of three subunits of 9-mers. Further, the minimum percentage difference between any two bar codes within this pool is 4/27 or 14.8%. The average pair-wise sequence difference in this pool of 1,000,000 genetic bar codes is 65%, or 17.5 nucleotides. In a preferred embodiment, to ensure hybridization stability, the number of G nucleotides in an MMS nucleotide subunit ranges from 45% to 50%. The number of G nucleotides is the same in all bar codes. A pool of 36-mer genetic bar codes, each constructed from four 9-mers of this MMS has a diversity of one hundred million. The minimal pair-wise sequence difference mismatch between any two 36-mers in this pool of

genetic bar codes is 14.8% while the average pair-wise sequence mismatch is 65%, or 23.4 nucleotides.

The following example lists one hundred 9-mers of a minimal mismatch set, each having four G nucleotides and five nucleotides selected from the group consisting of A, T, and an A plus T mixture, and further having a pair-wise mismatch of at least 4 nucleotides between any two 9-mers. Calculated using the formula set forth above, there exists 4,032 members in this MMS of 9-mers (*i.e.* $2^5C^4_9$). The typical number of sequences in each of these 4,032 sets ranges from 80 to 101. A pool of 27-mer genetic bar codes, each constructed from three 9-mers of this MMS, has a diversity of one million. The minimum pair-wise sequence mismatch between any two 27-mers in this pool of genetic bar codes is 14.8% while the average pair-wise sequence mismatch is 65%. Likewise, a population of 36-mer genetic bar codes, each constructed from four 9-mers of this MMS, has a diversity of one hundred million. The minimal pair-wise sequence mismatch between any two 36-mers in this pool is 14.8%, while the average pair-wise sequence mismatch is 65%, or 23.4 nucleotides.

The following one hundred 9-mers each has four G. This list sets forth one of the 4032 minimal mismatch sets available under this scheme.

GGGTATGAA
GGGGAAAAT
GGGGTATTA
GGGAGTATA
GGGAAGTTT
GGGTTTAGT
GGGATTTAG
GGAGGTTAA
GGAGAGATA
GGTGTGTAT
GGAGTTGTT
GGTGATTTG
GGAAGGAAT
GGTTGGTTA
GGTAGAGAA

GGATGAAGA
GGTAGTTGT
GGAAGATTG
GGTTGTAAG
GGAATGTGA
GGATAGTAG
GGTATGATG
GGAAAAGGT
GGTTTATGG
GAGGGTTTT
GAGGAGTAA
GTGGTGATT
GTGGATAGA
GTGTGGAAA
GAGTGAGAT
GAGAGATGA
GAGATGGTA
GAGTAGATG
GTGTTAGGA
GTGAAAGAG
GAAGGAGTA
GTTGGTGAT
GATGGAAGT
GTAGGAAAG
GATGAGGTT
GTAGTGGAA
GTAGAGTGT
GAAGTGTTG
GATGTTGGA
GAAGATGAG
GTTGTAGTG
GTATGGGTT

GATAGGTAG
GTAAGTGGA
GAATGTTGG
GAATAGGGA
GTTATGGGT
GTATTGAGG
GTTTATGGG
AGGGTGAAA
AGGGATTGT
TGGGTTATG
AGGTGGATT
TGGAGGTAA
AGGAGTGAT
TGGTGAGTA
TGGAGAAGT
AGGTGATAG
TGGTTGGAT
TGGTAGAGA
TGGATTGGA
AGGATAGTG
TGAGGGTTT
AGTGGAGTT
AGTGGTAGA
AGAGAGGAT
TGTGTGGTA
TGTGAGAAG
AGAGTAGGA
AGTGTTGAG
TGAGAAGTG
AGAAGGGTA
TGATGTGGT
TGTAGTGTG

AGTTAGGTG
AGAAAGAGG
ATGGGGTAT
TAGGGGATA
ATGGGTGTA
AAGGGTAAG
TTGGGATTG
TAGGTGTGT
TAGGAAGGA
ATGGTAAGG
TTGTGTGAG
ATGAGTTGG
AAGAAGGGT
AAGTTTGGG
AATGGGGAA
TTTGGGTGA
ATTGGGATG
AATGAGTGG
TTAGTTGGG
TATTGGAGG
AAAAGAGGG

If a restriction endonuclease or other mechanism is used to generate a single-stranded region in a bar-coded vector, then all four nucleotides may be used for design and synthesis of the genetic bar code. For example, restriction endonucleases such as BbvI, Bbsl, Bsal, BsmA 1, BsmF 1, BspM 1, FokI, Hga I and SfaN I may be used to generate an end having four or five protruding nucleotides.

5.2.7. PRODUCTION OF SINGLE-STRANDED GENETIC BAR CODES IN DOUBLE STRANDED LIBRARY VECTORS

An example restriction endonuclease for linearizing a bar-coded library and further exposing the genetic bar codes in single stranded form is Bgl II which may be used at site 1

in FIG. 5. Bgl II generates a 3' recessed end lacking G. T4 DNA polymerase is then used in the presence of GTP (and the absence of other NTPs) to digest the complementary strand of the genetic bar code. If an enzyme such as EcoR I is used for site 2 (FIG. 1), T4 DNA polymerase will stop degradation at the EcoR I site when it encounters nucleotide G. T4 DNA polymerase is then inactivated by heat. The bar-coded cDNA library having protruding single-stranded bar codes may then be purified using standard phenol/chloroform extraction and ethanol precipitation. For the convenience of cloning, a Bgl II site plus two additional nucleotides at its 5' end (for effective digestion with BgL II) is synthesized as a standard component of the vector proceeding the first nucleotide of the genetic bar code. Likewise, an EcoR I site plus one additional nucleotide at its 3' end (for effective digestion with EcoR I) is synthesized as a standard component of the vector after the last nucleotide of the genetic bar code.

5.2.8. HYBRIDIZATION OF A BAR-CODED LIBRARY WITH A GENE CHIP, WITH BAR-CODED BEADS OR WITH A BIOLOGICAL ARRAY

Hybridization of a bar-coded cDNA library with a genetic bar code population (whether on gene chips, beads or biological arrays) is carried out several hours to overnight at an optimal temperature, preferably 5 to 10 °C lower than the Tm or 2 to 5°C below Td of the genetic bar code population. The prehybridization buffer may be as follows: 6 x SSC (or 6 x SSPE), 0.01 M sodium phosphate (pH 6.8), 1 mM EDTA (pH 8.0), 0.5% SDS, 100 µg/ml denatured, fragmented salmon sperm DNA, and 0.1% nonfat dried milk. Hybridization buffer may be 3.0 M TMA Cl or 2.4 M TEA Cl, 0.01 M sodium phosphate (pH 6.8), 1 mM EDTA (pH 7.6), 0.5% SDS, 100 µg/ml denatured, fragmented salmon sperm DNA, and 0.1% nonfat dried milk. A hybridized bar-coded cDNA library may be washed with 6 x SSC solution and 2 x SSC solution as needed.

5.2.9. OTHER APPLICATIONS OF cDNA OR GENOMIC LIBRARIES SORTED USING GENETIC BAR CODES

A sorted, single-stranded cDNA library or genomic library can be used to create a "library array" for studying differential gene expression, as described below.

To make a single-stranded library array, a bar-coded library having a protruding single-stranded genetic bar code region is hybridized with a gene chip. This is followed by incubation with a DNA ligase, such as T4 ligase, to covalently bond the complementary strand of cDNA to the gene chip. The sense strand of cDNA may be removed by, for example, denaturing (*e.g.* 100 °C incubation) under high stringency or through using an enzyme such as Exo III and the like (Hoheisel, 1993, Anal. Biochem. 209:238), so as to convert a double stranded library into a single stranded library. For preparation of a single stranded antisense library, the genetic bar code region is placed in front of a CMV promoter and the same strand as sense cDNA. For preparation of a single stranded sense library, the genetic bar code region is placed at the end of the cDNA and the same strand as antisense cDNA.

Among the many advantages of making such a sorted library array using the genetic bar code technology of the invention are the following: 1) such a library array can be constructed at a density as high as an oligonucleotide array on a gene chip; 2) only a single sample is needed instead of a million samples for individual spotting for the current DNA array technology (*i.e.* since the cDNA library can be sorted using the genetic bar code method, only one sample containing the whole cDNA library is prepared, instead of preparing 10^6 individual samples as would be otherwise required); and 3) many different cDNA library arrays can be easily prepared (*i.e.* library arrays can be made from a plurality of bar coded cDNA libraries obtained from various sources without changing the format of the gene chip).

5.2.10. BEADS CAN REPLACE CHIPS WHEN USING GENETIC BAR CODES TO SORT A LIBRARY

In another embodiment of the invention, beads may be used instead of a gene chip or biological array, for sorting a cDNA library. In this embodiment, each individual bead carries multiple copies of one unique complementary bar code. Therefore, each individual bead will hybridize to multiple copies of a single recombinant, thereby sorting and concentrating individual members of the library to discrete loci. As spherical or spheroid supports which can migrate in solution, beads may provide enhanced hybridization efficiency compared to gene chips or biological arrays. Following hybridization, each bead represents a single, easily manipulable recombinant which may be assayed under a high-

throughput pharmaceutical screening format (*e.g.* one bead per well) to determine biological functions of the encoded cDNAs. For example, one can place each unique bead, with a bar-coded vector DNA hybridized thereto, into a well of an assay plate or into a micro-compartment of a micro-compartmentalization device (*e.g.* a 96-well plate, a 384-well plate, or a plate having a larger number of wells or micro-compartments). Such a micro-compartmentalization device may be as described in FIG. 1 through FIG. 4 of copending U.S. Patent Application No. 09/065,776, filed April 24, 1998, entitled "MICRO-COMPARTMENTALIZATION DEVICE AND USES THEREOF," by Cen and Sun (Attorney Docket No. 9557-003), which is incorporated herein by reference in its entirety. Bead placement may be performed robotically. The presence of a low salt solution in each well permits dissociation of vector DNA from the beads. The resulting solution in each well, now containing DNA of a single recombinant, may be mixed with a chemically transfected or electroporated into readout cells. A removable micro-compartmentalization device as described herein may be used during the transfection or electroporation procedure. If such a device is used, then the grid of the device may be removed from the readout cell culture following gene transfer so as to facilitate processing (*i.e.*, rinsing, culturing, and assaying for biological function). Any recombinant producing a positive signal in a readout assay may be recovered, for example, by sampling the positive cell population and performing PCR using primers flanking the cDNA insert of the vector.

5.2.11. AN ALTERNATIVE METHOD FOR GENERATING A SINGLE STRANDED GENETIC BAR CODE REGION ON A DOUBLE STRANDED VECTOR

To selectively expose single stranded DNA encoding the genetic bar code region of a double stranded vector, a protein binding site may be installed in the vector. A DNA binding protein which recognizes the protein binding site in the vector can then be used to sterically hinder (*i.e.* block or prevent) the 3' to 5' exonuclease progression beyond the bar code region of the vector. Such DNA binding proteins may include prokaryotic proteins such as a lactose repressor which binds to a lactose operator sequence, a tetracycline (tet) repressor which binds to a tet operator sequence, *etc.*, or eukaryotic proteins such as a eukaryotic transcription factor (*e.g.* E2F, AP1, SP1, p53, *etc.*). Any chosen protein binding site may be installed outside of the genetic bar code region defined by site 1 and site 2 using

standard recombinant DNA techniques (see FIG. 1). When cDNA library plasmids are linearized at site 1, DNA binding proteins may be applied to occupy two sets of DNA binding sites. Subsequently, a 3' to 5' single strand exonuclease (*e.g.* exonuclease III, T4 DNA polymerase, Klenow fragment, T7 DNA polymerase, Vent DNA polymerase, Pfu DNA polymerase, *etc.*) may be used to remove the complementary strand of the genetic bar code and site 2 for the genetic bar code end, and to remove site 1 for the non genetic bar code end. Exonuclease activity is stopped at the DNA binding sites by steric hindrance from the bound DNA binding proteins. Then the exonuclease is heat inactivated and both DNA binding proteins and exonuclease are removed from the DNA by phenol/chloroform extraction. When exonuclease III is used as the 3' to 5' exonuclease, the protection of the non genetic bar code end from 3' to 5' exonuclease activity can also be achieved by generating a 3' overhang which is resistant to exonuclease III. This approach may be used as an alternative to the DNA-protein complex formation described above which sterically hinders (*i.e.*, blocks or prevents) exonuclease progression along the DNA. A 3' overhang for this purpose can be obtained by installing a restriction enzyme site which produces such an overhang on the right side of site 1 (*see* FIG. 1). Suitable enzymes include Hae II, Kpn I, Nsi I, Pst I, Sac I, etc.

If a single stranded genetic bar code region is generated in the manner described above, all four nucleotides may be included in construction of the pool of unique genetic bar codes (as opposed to using only three nucleotides as described herein). For a given nucleotide building block length (*e.g.* 3-mers to 10-mers or more), the number of unique sequences in any minimal mismatch set (MMS) for a given mismatch cutoff will be larger. In other words, it is possible to achieve a higher percentage of minimal pair-wise mismatch when using all four nucleotides. Accordingly, one benefit to using all four nucleotides is to achieve a reduction of potential cross hybridizations among similar but non-identical bar codes. Of course, this benefit must be weighed against the benefits of using bar codes consisting of three nucleotides in a given situation.

5.2.12. ALTERNATIVES FOR DESIGNING GENETIC BAR CODES NOT USING AN OLIGONUCLEOTIDE BUILDING BLOCK APPROACH

Computer algorithms may be used to facilitate the design of unique genetic bar code sets having oligos of a given length, minimal cross hybridization, and the same or similar melting temperature (Tm) (*see e.g.* U.S. Patent Nos. 5,635,400 and 5,654,413 by Brenner).

Non-natural nucleotides (*i.e.* modified nucleotides or nucleotide derivatives) may be used when generating a complementary genetic bar code array to enhance the binding affinity between genetic bar codes and their complements.

Still further, when using the three nucleotide strategy to generate a protruding single stranded genetic bar code region from a double stranded vector, any DNA polymerase having 3' to 5' exonuclease activity may be used. Such polymerases include Klenow, T7, Vent, Pfu, and T4 DNA polymerases.

5.3. OUTPUT CONSIDERATIONS

Methods are provided to systematically screen expressed genes of the human genome for specific functions using a large number of functional assays. Such technology provides a very rapid system for gene identification in which at least some functional information may be inferred. Readout assays may be cell-based or biochemical-based. In examples of cell-based readout assays, changes in cellular morphology, immunostaining, or reporter gene expression can be detected within 1-2 days after library gene expression. In examples of biochemical-based readout assays, changes in ligand binding, growth factor activity, or enzymatic activity can be detected within 1-2 days after library gene *in vitro* transcription and translation. Using either approach, full functional screening of a library having a diversity of 10^6 can be completed within a few days.

In this way, the methods of the invention provide the advantages of minimizing workload and reducing the occurrence of gene mutations which can arise in screening assays employing long-term culture. It is estimated that one of ordinary skill in the art can easily screen one library per week by the methods of the invention without using automation. Of course, if automation is used, multiple libraries per week may be screened.

Cell based immunostaining assays have been extensively used under the single-gene approach for detection of gene expression, subcellular localization and biological functions.

Two examples of established mammalian cell-based immunostaining assays are as follows. First, cytochemical detection of intracellular LDL-derived cholesterol accumulation has been used to demonstrate that cholesterol accumulates in fibroblasts of individuals having Niemann-Pick C1 disease (*see* Carstea et al., Science 277, 228-231). Second, cellular localization of heat shock transcription factor (HSF) has been used to demonstrate that HSF nuclear localization changes from a uniform distribution to a punctate distribution when staining for activated HSF after c-myb expression in 293T cells (*see* Kanei-Ishii et al., Science 277, 246-248).

Similar to these two examples, functional assays of use together with the methods of the invention will measure changes (*i.e.* induction or reduction) of target gene expression, changes of cellular localization of a specific antigen, changes in cellular behaviors (*e.g.* growth factor secretion, apoptosis factor secretion, differentiation factor secretion), and changes in cellular morphology.

There are many assays of gene function in existence, each having a particular readout. For example, induction or reduction of target gene mRNA or protein expression can be detected by means standard in the art, including nucleic acid hybridization and antibody detection of specific antigens.

There are at least three categories of readout assays for use with the methods of the invention: (a) assays for genes associated with signaling pathways; (b) assays for genes associated with specific diseases; and (c) assays for genes associated with cellular physiological functions. Each of these categories is further described below.

5.3.1. PATHWAYS

The signals a cell receives, whether from outside or inside the cell, are generally transmitted through a cascade of molecular interactions, including protein-protein interactions. The overall process is generally termed signal transduction. The signaling pathways which may be assayed for identification of associated genes include, but are not limited to, a mitogenic signaling pathway, a STAT signaling pathway, an NFκB signaling pathway, a stress signaling pathway, an apoptosis signaling pathway, an NFAT signaling pathway, a Wnt signaling pathway, a CREB signaling pathway, an AP-1 signaling pathway, a proliferation signaling pathway and an anti-proliferation signaling pathway.

For proliferation signaling, BRDU incorporation or PCNA induction can be the cellular event detected. For stress signaling, p53 induction, Jun induction, or nuclear HSF3 aggregates can be the detectable cellular event. For apoptosis signaling, detection by ApoAlert™ staining (Gavrieli *et al.*, 1992, J. Cell. Biol. 119, 493) or annexin staining can be the detectable cellular event. For anti-proliferation signaling, detection of p21, p27, p57, p15, p16, p18, or p19 induction can be the detectable cellular event. For Wnt signaling, detection of β-catenin induction or β-catenin re-localization can be the detectable cellular event. For STAT signaling, detection of induction of a reporter gene under the control of a STAT1, 2, 3, 4, 5, 6, or 7 promoter can be the detectable cellular event. For AP-1 signaling, detection of c-fos induction can be the detectable cellular event. For CREB signaling, CREB phosphorylation, or induction of a reporter gene under the control of the CREB promoter, can be the detectable cellular event. For NFκB signaling, NFκB re-localization, or induction of a reporter gene under NFκB promoter control, can be the detectable cellular event. For NFAT signaling, IL-2 mediated proliferation can be the detectable cellular event. Other signaling pathways can include Hedgehog signaling (detectable by GLI-1, GLI-2, and GLI-3 induction); nuclear receptor signaling (detectable by induction or reduction of a reporter gene under estrogen, retinoic acid, vitamin D3 or thyroid hormone responsive promoters); antiviral signaling (detectable by induction of interferon alpha or beta); myc-max signaling (detectable by induction of a reporter gene under a myc-Max responsive promoter); BMP signaling (detectable by nuclear translocation of Smad); and insulin signaling (detectable by Glut1 or Glut4 re-localization).

5.3.2. DISEASES

Specific diseases of interest include, but are not limited to, cancer, inflammation, atherosclerosis, autoimmune diseases, diabetes, infection, diseases of metabolism (*e.g.* obesity), and neurodegenerative diseases (*e.g.* Alzheimer's disease and Parkinson's disease). Readout assays involving detection of changes (*i.e.* increases or decreases) in the levels of the following targets may identify genes associated with the indicated specific disease.

Assays that may detect genes involved in cancer include assays for detection of: HLA for immune surveillance; OSM for anti-cancer growth; GADD45 and GADD153 for tumor suppression; nm23 for anti-metastasis; vEGFA, vEGFB, vEGFC, PIGF, and FGF2 for angiogenesis; MDR for drug resistance; CASP100 for apoptosis; and PDGFA, PDGFB,

FGF1, 3, 4, 5, 6, 7, 8, and 9, IGF 1, IGF 11, cyclin A, B1, C, D1, D2, D3, E, F, G1, and H, c-myc and c-Jun for growth. Assays that may detect genes involved in inflammation include detection of Cox2, IL-1β, IL-6, TNFα, IL-13, E-selectin, VCAMI, ICAM 1 and 2, NFκB, c-Rel, RelB, IκBα, IκBβ, and Bcl3.

Changes in the level of expression of the following targets may be assayed immunologically in response to expression of a heterologous cDNA in order to detect genes which may be involved in a given disease. For example, the potential targets that can be used for detecting genes involved in atherosclerosis include Egr-I. The potential targets that can be used for detecting genes involved in autoimmunity include Fas and Fas ligand. The potential targets that can be used for detecting genes involved in diabetes include insulin. The potential targets that can be used for detecting genes involved in infection include chemokines (MIP-1α, MIP-1β, MIP-2, RANTES, MCP-1, MCP-2, GROα, GROβ, GROγ, ENA-78, 1309, and IP10) and various cytokines (*e.g.* IL-2, IL-13, GM-CSF, G-CSF, and M-CSF). The potential targets that can be used for detecting genes involved in obesity include leptin and the leptin receptor. The potential targets that can be used for detecting genes involved in Alzheimer's disease include Tau, CRF, CRF receptor, CRF-BP, Urocortin, and neuronal growth factors (*e.g.* BDNF, NT3, NT4, NT5, CNTF, and GDNF). The potential targets that can be used for detecting genes involved in Parkinson's disease includes TH, and α-synuclein.

5.3.3. FUNCTIONS

Where identification of genes associated with various physiological functions is desired, an assay may be employed which can detect changes in such functions as cell growth, apoptosis, senescence, differentiation, adhesion, binding of a cell to a specific molecule, binding of a cell to another cell, cellular organization, organogenesis, intracellular transport, transport facilitation, protein synthesis, transcription, energy conversion, metabolism, myogenesis, neurogenesis, or hematopoiesis. Examples of such cellular physiological functions and assays for detecting changes in them include, but are not limited to: cholesterol transport, detectable by detecting intracellular cholesterol accumulation; myogenesis, detectable by detecting induction of MyoD or MEF-2; neurogenesis, detectable by detecting induction of neuro D; and vasodilation and neurotransmission, detectable by induction of inducible nitric oxide synthase (iNOS).

A number of specific exemplary assays which may be used to identify genes in conjunction with the methods of the invention are set forth in detail below.

5.4. CELLULAR READOUT ASSAYS

5.4.1. PROLIFERATION PATHWAY

Bromodeoxyuridine (BRDU) incorporation may be used as an assay to identify genes involved in proliferation. The BRDU assay identifies a cell population undergoing DNA synthesis by incorporation of BRDU into newly-synthesized DNA. Newly-synthesized DNA may then be detected using an anti-BRDU antibody (*see* Hoshino *et al.*, 1986, Int. J. Cancer 38, 369; Campana *et al.*, 1988, J. Immunol. Meth. 107, 79).

A proliferating cell nuclear antigen (PCNA) assay may also be used to identify genes involved in cell proliferation. PCNA (*a.k.a.* cyclin or the polymerase d associated protein) is a 36 kilodalton protein whose expression is elevated in proliferating cells. PCNA is synthesized in early G1 and S phases of the cell cycle and therefore serves as an excellent marker for proliferating cells. Positive cells are identified by immunostaining using an anti-PCNA antibody (*see* Li *et al.*, 1996, Current Biology 6, 189; Vassilev *et al.*, 1995, J. Cell Sci. 108, 1205).

5.4.2. STRESS SIGNALING PATHWAY

p53 is an important modulator of the stress response. p53-dependent transcriptional activation may therefore be used to identify genes involved in a stress signaling pathway. A readout cell population containing a reporter gene under the control of a p53-inducible promoter may be used for the assay. Suitable reporter genes include, but are not limited to, β-galactosidase (β-gal), chloramphenicol acetyltransferase (CAT), and luciferase. Positive cells may be identified by blue color in a β-gal reporter gene assay (*see e.g.* Komarova *et al.*, 1997, EMBO J. 16, 1391-1400) or by immunostaining for the reporter gene product. A p53 induction assay may also be used to identify genes involved in a stress signaling pathway. p53 induction (*i.e.* increases in cellular p53 protein expression) may be identified by immunostaining using a specific anti-p53 antibody (Anker *et al.*, 1993, Int. J. Cancer 55, 982; Weiss *et al.*, 1993, Int. J. Cancer 54, 693).

A heat shock transcription factor 3 (HSF3) aggregation assay may also be used to identify genes in a stress signaling pathway. The HSF3 aggregation assay measures HSF3

aggregation in the nucleus induced by cellular stress signals through immunostaining using a specific anti-HSF3 antibody (Kanei-Ishii *et al.*, 1997, Science 277, 246).

An activated c-Jun kinase assay may be used to identify genes in a stress signaling pathway. c-Jun kinase (JNK) is a transcription factor which is activated by phosphorylation (p-JNK). Many stress signals result in activation of c-Jun kinase by phosphorylation (Derijard *et al.*, 1994, Cell 76, 1025). The availability of p-JNK specific antibodies (Santa Cruz) allows *in situ* detection of cells in which JNK is activated by heterologous library genes.

5.4.3. LOSS OF INVASIVENESS

Invasion inhibition assays may be used to identify genes involved in cancer. One such assay measures induction of E-cadherin-mediated cell-cell adhesion. The induction of E-cadherin-mediated adhesion can result in phenotypic reversion and loss of invasiveness of epithelial cells. This assay measures increased expression of E-cadherin at the cell junction through immunostaining using a specific anti-E-cadherin antibody (Hordijk *et al.*, 1997, Science 278, 1464). Another such assay measures loss of hepatocyte growth factor (HGF)-induced cell scattering. Loss of HGF-induced cell scattering is correlated with loss of invasiveness of epithelial cells such as Madin-Darby canine kidney (MDCK) cells. This assay identifies a cell population which has lost cell scattering activity in response to HGF and therefore forms compact colonies (Hordijk *et al.*, 1997, Science 278, 1464).

5.4.4. APOPTOSIS SIGNALING PATHWAY

One assay for apoptosis is the terminal deoxynucleotidyl transferase-mediated dUTP nick-end-labeling (TUNEL) assay. The TUNEL assay is used to measure nuclear DNA fragmentation, the hallmark of apoptosis in many cell types (*see e.g.* Lazebnik *et al.*, 1994, Nature 371, 346), by following the incorporation of fluorescein-dUTP (Yonehara *et al.*, 1989, J. Exp. Med. 169, 1747). These assay kits are commercially available through suppliers such as Clontech and Boehringer Mannheim.

5.4.5. ANTI-PROLIFERATION PATHWAY

One assay useful for gene identification in an anti-proliferation signaling pathway is the p15 induction assay. p15 is a member of a family of specific inhibitors of Cdk4 and

Cdk6. The latter are essential for G1 progression into S phase of the cell cycle (Sherr *et al.*, 1995, Genes & Dev. 9, 1149). The expression of p15 is positively regulated by transforming growth factor-β (Reynisdottir *et al.*, 1997, Genes & Dev. 11, 492). p15 induction may be identified by immunostaining using a specific anti-p15 antibody available commercially (*e.g.* Santa Cruz).

Another assay useful for gene identification in an anti-proliferation signaling pathway is the p21 induction assay. Increased levels of p21 expression in cells results in Cdk inhibition, thus resulting in delayed entry into G1 of the cell cycle (Harper *et al.*, 1993, Cell 75, 805; Li *et al.*, 1996, Current Biology 6, 189). For example, p21 expression can be elevated by p53 and transforming growth factor-β activities. p21 induction may be identified by immunostaining using a specific anti-p21 antibody available commercially (*e.g.* Santa Cruz).

Yet another assay useful for gene identification in an anti-proliferation signaling pathway is the p27 induction assay. As for the assays above, p27 is also a member of the Cdk inhibitor family of proteins. The expression of p27 is increased upon mitogen withdrawal or contact inhibition (Polyak *et al.*, 1994, Cell 78, 59). p27 induction may be identified by immunostaining using a specific anti-p27 antibody available commercially (*e.g.* Santa Cruz).

5.4.6. WNT SIGNALING PATHWAY

One assay for detection of genes which modulate the Wnt signaling pathway is a β-catenin induction and/or translocation assay. The activation of the Wnt signaling pathway results in an increased expression of β-catenin and the translocation of β-catenin from the cytoplasmic compartment to the nucleus (Kuhl *et al.*, 1997, BioEssays 19, 101). This assay is used to identify cells and/or cell populations in which the expression of β-catenin is increased compared to background levels, and/or in which a change of β-catenin localization occurs, in response to expression of a heterologous gene. Changes in β-catenin expression or localization are detected using a specific anti-β-catenin antibody (*e.g.* Tao *et al.*, 1996, J. Cell Biol. 134, 1271).

Another assay for detection of genes which modulate the Wnt signaling pathway is a LEF-1 inducible promoter induction assay. β-catenin activates downstream targets in the Wnt signaling pathway by binding to a transcription factor known as LEF-1, thus resulting

in activation of a LEF-1 inducible promoter (Korinek *et al.*, 1997, Science 275, 1785). A readout cell line containing a reporter gene, such as β-gal, under a LEF-1 inducible promoter is used for the assay. When β-gal is used as a reporter gene, positive cells are the darker blue cells.

5.4.7. STAT SIGNALING PATHWAY

The STAT (signal transducers and activators of transcription) signaling pathway is activated by many growth factors and cytokines and plays essential roles in cell differentiation, cell cycle control, and development. There are six known members of the STAT transcription factor family. Each STAT family member (except STAT2) is known to recognize a specific DNA binding sequence (Ihle, 1996, Cell 84, 331). The assay employs a readout cell line containing a reporter gene, such as β-gal, under the control of any of these known STAT-inducible promoters (White *et al.*, 1996, Cytokine Growth Factor Rev. 7, 303). Positive cells stain dark blue when β-gal is used as the reporter gene. This assay may be used to identify genes in a STAT1 signaling pathway, a STAT3 signaling pathway, a STAT4 signaling pathway, and/or a STAT5/STAT6 signaling pathway. Since STAT5 and STAT6 share the same DNA recognition site, the assay does not distinguish between these two STAT pathways. Readout cells expressing a gene which activates a particular STAT transcription factor will produce a positive signal. Accordingly, the genes identified reside just upstream in the particular STAT pathway assayed.

5.4.8. MAP KINASE SIGNALING PATHWAY

MAP kinase signaling pathway genes may be identified using a p-ERK assay. The activation of this signal transduction pathway by certain growth factors, hormones and neurotransmitters is mediated through two closely-related MAP kinases, p44 and p42, also known as ERK1 and ERK2. ERK proteins are activated by dual phosphorylation at specific tyrosine and threonine sites. The p-ERK assay is used to identify genes by immunostaining readout cells with an antibody which specifically detects the presence of phosphorylated ERK (p-ERK). Such p-ERK antibodies, which only recognize phosphorylated ERK1 and ERK2, may be obtained commercially (*e.g.* Santa Cruz). *See* Boulton *et al.*, 1991, Cell 65, 663.

5.4.9. AP-1 SIGNALING PATHWAY

Genes in an AP-1 signaling pathway may be identified using a c-fos induction readout assay. The AP-1 signaling pathway is involved in cell proliferation, cell survival and cell stress. Activation of the AP-1 signaling pathway results in an increased expression of genes under the control of an AP-1 promoter sequence such as the c-fos gene (*see e.g.* Karin *et al.*, 1997, Curr. Opin. Cell Biol. 9, 240). The c-fos induction assay identifies genes expressed in cell populations in which the level of endogenous c-fos protein is increased by immunostaining c-fos using a specific anti-c-fos antibody (Telford *et al.*, 1996, J. Comp. Neurol. 375, 601).

5.4.10. CREB SIGNALING PATHWAY

In one embodiment, genes in a cyclic adenosine monophosphate response element binding protein (CREB) signaling pathway may be identified using a phosphorylated CREB (p-CREB) readout assay. CREB is activated by phosphorylation following an increase in the intracellular concentration of cAMP or Ca^{2+}. An antibody which specifically recognizes phosphorylated CREB allows detection of an activated CREB pathway in readout cells (Ginty *et al.*, 1994, Cell 77, 713).

In another embodiment, genes in a CREB signaling pathway may be identified using a cyclic adenosine monophosphate response element (CRE) reporter gene assay. In this assay, a readout cell containing a reporter gene (*e.g.* β-gal, CAT or luciferase) under the control of the CRE is used for the assay. Positive cells may be identified by, *e.g.*, blue staining in a β-gal assay (Himmler *et al.*, 1993, J. Recept. Res. 13, 79; Kruger *et al.*, 1997, Naunyn Schmiedebergs Arch. Pharmacol. 356, 433) or by immunostaining for the reporter gene product.

5.4.11. NFκB SIGNALING PATHWAY

In one embodiment, an NFκB translocation assay may be used to identify genes in an NFκB signaling pathway. Activation of the NFκB signaling pathway results in translocation of NFκB from the cytoplasm to the nucleus. The NFκB translocation assay identifies cells with NFκB translocated to the nucleus by immunostaining for NFκB using a specific anti-NFκB antibody (Han *et al.*, 1997, J. Biol. Chem. 272, 9825; Janssen *et al.*, 1995, Adv. Cancer Res. 151, 389).

In another embodiment, an NFκB reporter gene assay may be used to identify genes in an NFκB signaling pathway. In this assay, a readout cell containing a reporter gene (*e.g.* β-gal, CAT or luciferase) under the control of an NFκB response element is used for the assay. Positive cells may be identified, *e.g.*, by blue staining in a β-gal assay or by immunostaining for the reporter gene product (Rothe *et al.*, 1995, Science 269, 1424).

5.4.12. NFAT SIGNALING PATHWAY

Genes in an NFAT signaling pathway may be identified using a NFAT reporter gene assay. In this assay, a readout cell expressing a reporter gene (*e.g.* β-gal, CAT or luciferase) under the control of an NFAT response element is used. Positive clones may be identified by blue staining in a β-gal assay (*see e.g.* Burres *et al.*, 1995, J. Antibiot. 48, 380) or by immunostaining for the reporter gene product.

5.4.13. INSULIN SIGNALING PATHWAY

Genes in the insulin signaling pathway may be identified using a GLU4 translocation assay. Insulin stimulation of adipocytes results in translocation of the GLU4 glucose transporter to the plasma membrane. This assay identifies cells in which the insulin signaling pathway is activated by immunostaining GLU4 protein localized at the plasma membrane (Martin *et al.*, 1996, J. Biol. Chem. 271, 17605).

5.4.14. MDR SIGNALING PATHWAY

Genes in the multiple drug resistance (MDR) gene regulation pathway may be identified using an MDR reporter gene assay. MDR gene expression is often greatly increased in cancer cells resistant to chemotherapy. In this assay, a readout cell containing a reporter gene (*e.g.* β-gal, CAT or luciferase) under the control of an MDR gene promoter may be used for the assay. Positive cells may be identified by blue staining in a β-gal assay (Walther *et al.*, 1997, Gene Ther. 4, 544) or by immunostaining using an antibody specific for the reporter gene product.

5.4.15. CHOLESTEROL TRANSPORT PATHWAY

Genes important in a cholesterol transport pathway may be identified using an intracellular cholesterol accumulation assay. For example, mutations of the Niemann-Pick

type C (NP-C) gene result in lysosomal accumulation of low density lipoprotein (LDL)-derived cholesterol. The accumulated cholesterol in the cytoplasm is detected by staining with filipin, a specific cytochemical marker of unesterified cholesterol. The filipin staining assay may be used to identify cells with cholesterol accumulation due to the expression an exogenous sense or anti-sense cDNA (*see* Eugene *et al.*, 1997, Science 277, 228).

5.5. BIOCHEMICAL READOUT ASSAYS

In the practice of this invention, biochemical readout assays may be used to identify genes modifying specific activities following *in vitro* transcription and translation. Such biochemical readout assays include, but are not limited to, enzymatic and receptor-based assays. There are a wide variety of assays for enzymatic activities and receptor-binding activities which may be adapted for use in identification of new genes upon screening a library of interest, as further exemplified in this Section below.

There are many resources available describing such enzymatic and receptor-based assays suitable for use with the methods of the invention. For example, *Methods in Enzymology* is a multi-volume reference published by Academic Press which describes biological methods, including enzymatic and receptor-based assays, in detail. Further, Fernandez-Botran and Vetvicka, 1995, *Methods in Cellular Immunology*, CRC Press, describes assays for immune cell activation, including cytokine receptor assays.

Biochemical readout assays may include, *e.g.*, detection of: GABA receptor activity, glutamate receptor activity, monoamine oxidase activity, nitric oxide synthetase activity, opiate receptor activity, serotonin receptor activity, adenosine A_1 agonist and antagonist activity, adrenergic α_1, α_2, β_1 agonist and antagonist activity, calcium channel blocker activity, inflammatory mediator activity, such as the interleukins (*e.g.* IL-1, IL-6), tumor necrosis factor activity, arachidonic acid activity and phosphatase activity (*e.g.* tyrosine phosphatase). Further, biochemical readout assays may include, for example, binding to protein domain or subdomain, for example, a PDZ domain, a PH domain, an SH2 domain, and an SH3 domain. Still further, biochemical readout assays may include binding to a molecule, for example, phosphotyrosine and phosphorylated inositol. A functional assignment given to a particular gene may be derived from results obtained in more than one assay. Indeed, it is preferred that a functional assignment be derived from results

obtained in a panel of two or more assays. Generally, one skilled in the art would know which assays are appropriate to employ to best identify genes having, or modifying, a particular function-of-interest.

Further specific examples of assays based on enzymes or receptors include the following: acetylcholinesterase; aldol-reductase; angiotensin converting enzyme (ACE); cyclooxygenases; DNA repair; β-glucuronidase; lipoxygenases; monoamine oxidases; phospholipase A_2, platelet activating factor (PAF); potassium channel assays; prostaglandin synthetase; serotonin re-uptake activity; and steroid receptors. Additional assays may include: ATPase inhibition, benzopyrene hydroxylase inhibition, HMG-CoA reductase inhibition, phosphodiesterase inhibition, protease inhibition, and tyrosine kinase inhibition.

5.6. USER-DEFINED ASSAYS

The methods of the invention are not limited to use with the readout assays described herein. Such readout assays merely serve to exemplify a few of the myriad possibilities suitable for use with the invention. When the readout assay is a cellular readout assay, virtually any cell line identified as suitable by one skilled in the art may be used. Further, virtually any reporter gene, or endogenous gene functioning as a reporter gene, identified as suitable by one skilled in the art may be used. It will be well noted by one skilled in the art that the methods of the invention are suitable for use with any known readout assay, whether the assay be cellular or biochemical.

The skilled practitioner will recognize that it is the particular readout assay, whether chosen from the literature or designed by the user, which determines the type (*i.e.* function) of genes identified. For example, if one wishes to identify genes associated with cancer, one may choose to screen the library of interest using the p53 and or MDR assays described above. Often, the user will provide the most appropriate readout assay to be employed for identification of particular genes-of-interest.

5.7. AUTOMATION

It is preferred that automation technology be applied throughout the entire functional gene identification process. Many steps in the overall process are amenable to such automation. For example, robotic colony picking may be used for building a library of 10^6 clones from plates containing well-isolated colonies. Robots suitable for this purpose are available commercially from, *e.g.*, Qiagen, Gentix, etc. Similarly, transfection of retroviral vectors into producer cells, and *in situ* transfection of bar-coded, sorted libraries into readout cells, are repetitive operations suitable for robotic automation. Further, the system is suitable for automated immunostaining of the co-culture, and to automated microscopic viewing of the immunostained result. Only one population of bar codes is needed for all screenings and the same nucleic acid array can be used repeatedly. Automation can be applied to hybridization to an array such that the same hybridization conditions are used for various libraries. Automation can also be applied to in situ transfections and in situ bioassays.

6. EXAMPLES

The following examples are provided merely as illustrative of several embodiments and should not be construed to limit in any way the invention.

6.1. ASSAYS FOR CELLULAR PROLIFERATION

In the proliferation/anti-proliferation assays described in this example, the function of genes identified will depend on the type of library screened. For example, if a sense cDNA library is screened, genes associated with proliferation will be identified. By contrast, if an antisense cDNA library is screened, genes associated with anti-proliferation will be identified.

In one assay, PCNA immunostaining is performed using readout cells that are starved for at least 24 hours in low serum (*e.g.* 0.5%) medium prior to tetracycline induction. After 12-24 hours of tetracycline treatment, the cells are fixed in cold methanol (*e.g.* 5 minutes at -20°C) and air dried. Cells are blocked with 10% normal goat serum in phosphate buffered saline (PBS) for 30 minutes at 37°C. This is followed by incubation with a 1:10 dilution of anti-PCNA antibody (*e.g.* PC10 antibody, Dakkopat) for 2 hours at 37°C. Cells are rinsed with PBS (*e.g.* three times for 5 minutes each) and incubated with

goat anti-mouse IgG antibody conjugated with fluorescein isothiocyanate (FITC; NBL, Nagoya) for 45 minutes at 37°C. After washing, cells are mounted in mounting medium (*e.g.* 3% hydroquinone, 50% glycerin, pH 8-9). Observation (*i.e.* readout visualization) is performed using a fluorescence microscope (Zeiss, Germany).

In another assay, BRDU incorporation is performed using readout cells similarly starved as above. 10 μM 5-bromo-2'-deoxyuridine (BRDU) may be added at the time of tet induction, or a few hours later (*e.g.* 2-12 hours). After 12-24 hours of induction, the readout cells are fixed (*e.g.* absolute methanol for 10 minutes at 4°C) and air dried. Cells are next rehydrated (*e.g.* PBS for 3 minutes), DNA is denatured (*e.g.* 2 M HCl for 60 minutes at 37°C), the preparation is neutralized (*e.g.* 0.1 M borate buffer, pH 8.5, 2X for 5 minutes) and washed (*e.g.* PBS 3X over 10 minutes). The readout cell preparation is incubated with anti-BRDU antibody (*e.g.* 50 μg/ml for 60 minutes at room temperature; Boehringer Mannheim), washed (*e.g.* PBS 3X over 10 minutes), and counterstained with Harris-modified hematoxylin. The preparation is then dehydrated and mounted for observation (*i.e.* visualization of readout cells staining positive with anti-BRDU antibody).

6.2. ASSAYS FOR p53 REGULATION

An assay for p53 induction may be used to identify genes associated with p53 expression. One or more p53 inducer genes may be identified if a sense cDNA library is screened. On the other hand, one or more p53 inhibitor genes may be identified if an antisense library is used. The p53 assays may be conducted by measuring endogenous p53 levels or by measuring levels of a reporter gene operably linked to a p53 promoter, as described below.

For an assay for p53 induction in which endogenous p53 expression is activated, readout cells are treated with tetracycline for 12-24 hours, to induce library gene expression, prior to anti-p53 antibody staining for endogenous p53 expression. Cells are fixed with 1:1 acetone:methanol at -20°C for 10 minutes and air dried. This is followed by blocking with 3% BSA in PBS for 30 minutes. Monoclonal antibody Pab 1801 (Novocastra, Newcastle, UK) which recognizes both wild-type and mutant p53, may be used at a dilution of 1:25 for a 1 hour incubation. After washing with PBS three times over 10 minutes, FITC conjugated anti-mouse IgG antibody (Cappel) may be used for detection.

For an assay of p53 induction using a reporter gene, a p53 promoter operably linked to a β-gal reporter gene may be used. Readout cells containing the β-gal gene under the direction of the p53 promoter can be obtained, *e.g.*, from transgenic mice (*see* Komarova *et al.*, 1997, EMBO J. 16, 1391-1400) or by establishing a stable cell line expressing a reporter gene under control of the p53 promoter. Such readout cells are induced with tetracycline, fixed with 1% glutaraldehyde in PBS, washed three times with PBS, and stained in 0.2% X-gal, 3.3 mM $K_4Fe(CN)_6$, 3.3 mM $K_3Fe(CN)_6$ and 1 mM $MgCl_2$ for one or more hours. Positive cells are detected by the characteristic blue color which develops from the β-gal staining.

6.3. HSF INTRACELLULAR TRANSLOCATION ASSAY

A heat shock transcription factor (HSF) intracellular translocation assay may be used to identify genes which are associated with transport of HSF from the cytoplasm to the nucleus. One or more HSF transport inducer genes may be identified if a sense library is screened. Alternatively, one or more genes associated with inhibition of HSF transport may be identified if an antisense library is screened. Readout cells may be fixed with either absolute methanol or 4% paraformaldehyde. After blocking with 10% normal goat serum in PBS, cells are incubated with a 1:300 dilution of anti-HSF3 in 10% normal goat serum in PBS, followed by FITC-conjugated goat anti-rabbit IgG antibody (1:200 dilution) (Cappel). The preparation may be washed and mounted as described above.

6.4. CHOLESTEROL TRANSPORT ASSAY

A filipin staining assay may be used to identify genes associated with blocking cholesterol transport when screening a sense cDNA library or to identify genes associated with facilitating cholesterol transport when screening an antisense cDNA library. The assay is based on the principle that filipin can specifically stain unesterified cholesterol located inside of cells (Carstea *et al.*, 1997, Science 277, 228). The presence of large amounts unesterified cholesterol inside of cells indicates breakdown of the cholesterol transport pathway.

After tetracycline induction, readout cells are washed three times with Dulbecco's PBS and fixed with 10% phosphate-buffered (pH 7.4) formalin at room temperature for a minimum of 1 hour. Cells are rinsed three times with Dulbecco's PBS before they are

stained in filipin (Sigma) for 60 minutes. The filipin staining solution is prepared by dissolving 2.5 mg of filipin in 1 ml of DMSO, which is then added to 50 ml of Dulbecco's PBS. The stained cells are washed three times with Dulbecco's PBS and mounted with glycerol/gelatin containing 1% phenol.

6.5. LIBRARY MUTATIONAL ANALYSIS OF A SINGLE GENE

This invention provides a high throughput method for structure-function analysis of a particular protein using random mutagenesis of a single gene of interest and the functional screening methods described herein. In this embodiment, a library containing randomly mutagenized recombinants from a gene are obtained using, *e.g.*, PCR with two primers framing the DNA region to be mutagenized under conditions of reduced Taq polymerase fidelity (*see e.g.* Rice *et al.*, 1992, Proc. Natl. Acad. Sci. U.S.A. 89:5467; Leung *et al.*, 1989, Technique 1:11). The mutagenized library may also be a deletion library which can be obtained through inverse PCR using 5'-truncated primers (Pues *et al.*, 1997, Nucl. Acids Res. 25:1303).

6.6. CONSTRUCTION OF A BIOLOGICAL ARRAY VECTOR

In one embodiment, when a bar-coded cDNA library is in double stranded form (except for the genetic bar code region), there is no limitation on the type of vector which may be used to construct the biological arrays. However, when a bar-coded cDNA library is in single stranded form, only vectors which share no homology with the cDNA library vector should be used to preclude vector hybridization outside of the bar code region. For example, if the cDNA library plasmid is pBR322 and contains an ampicillin resistance gene, suitable vectors for construction of biological arrays include plasmid M13 or pACYC (available from New England BioLabs) with deletion of ampicillin resistance gene. If the microbial used to construct a biological array is yeast, a yeast vector such as Yep24, Yip5, *etc.* (available from New England BioLabs) may be used.

6.7. MANUALLY-SORTED cDNAs AND OLIGONUCLEOTIDES FOR USE IN *IN SITU* TRANSFECTION AND CELLULAR READOUT ASSAYS

Described hereinabove is the use of nucleic acid arrays for sorting bar-coded cDNA libraries. In another embodiment, this invention provides a method for manually sorting cDNAs and oligonucleotides for screening using the *in situ* transfection procedures and cellular readout assays described herein. Such manual sorting has the advantage of not requiring a bar-coded vector. Manual sorting may be carried out by mechanical spotting of individual cDNAs onto a solid support (*e.g.* nitrocellulose or nylon). Such a manually-sorted cDNA population can be considered to be another form of a nucleic acid array. Such an array can also be used in the *in situ* transfection and cellular readout assays described herein, so long as the cDNA is cloned into an expression vector which is capable of expressing either sense or antisense cDNA (as desired by the user) when transfected into readout cells. In this way, a manually-sorted nucleic acid array can be used to analyze a collection of full-length genes-of-interest from any given source.

A manually-sorted single-stranded oligonucleotide array can also be constructed and used for the *in situ* transfection procedures and cellular readout assays described herein to identify a particular oligonucleotide which is most effective in manifesting a biological function-of-interest (*e.g.* antisense inhibition of oncogene expression). Such a manually-sorted oligonucleotide array may be obtained through mechanical spotting of individual oligonucleotides onto a solid support (*e.g.* nitrocellulose or nylon). Such an approach may be an effective way for identifying an optimal antisense oligonucleotide from among a population of antisense oligonucleotides which is effective in altering the expression of a particular target gene, such as the *ras* oncogene.

The invention described and claimed herein is not to be limited in scope by the specific embodiments herein disclosed since these embodiments are intended as illustration of several aspects of the invention. Any equivalent embodiments are intended to be within the scope of this invention. Indeed, various modifications of the invention in addition to those shown and described herein will become apparent to those skilled in the art from the foregoing description. Such modifications are also intended to fall within the scope of the appended claims. Throughout this application various publications and patents are cited. Their contents are hereby incorporated by reference into the present application in their entireties.

We claim:

1. A method for conducting a biological readout assay used to screen a bar-coded cDNA library comprising:

(a) sorting the bar-coded cDNA library using a nucleic acid array;

(b) transfecting the library sorted in step (a) into a readout cell line *in situ*; and

(c) conducting the biological readout assay.

2. The method of Claim 1, wherein the nucleic acid array is a biological array or a gene chip.

3. The method of Claim 2, wherein the biological array comprises a population of vectors, each vector containing a different bar code complementary to a bar code of the cDNA library to form a population of complementary bar codes, wherein the population of vectors is immobilized on a support.

4. The method of Claim 3, wherein the population of complementary bar codes consists of from 10^2 to 10^8 complementary bar codes.

5. The method of Claim 3, wherein the support is formed of nitrocellulose or nylon.

6. The method of Claim 1, wherein transfecting *in situ* is carried out using a chemical transfectant or electroporation.

7. The method of Claim 1, wherein the readout cell line is NIH 3T3 cells carrying a reporter gene under the control of a response element or promoter.

8. The method of Claim 7, wherein the reporter gene is selected from the group consisting of β-galactosidase, luciferase and chloramphenicol acetyltransferase.

9. The method of Claim 7, wherein the response element or promoter is selected from the group consisting of an NFκB response element, an NFAT response element, a cyclic adenosine monophosphate response element, a STAT-inducible promoter, a LEF-1-inducible promoter and a p53-inducible promoter.

10. The method of Claim 1, wherein expression of the bar-coded cDNA library is tetracycline inducible or estrogen inducible.

11. The method of Claim 1, wherein the biological readout assay is capable of detecting genes in a pathway selected from the group consisting of a mitogenic signaling pathway, a STAT signaling pathway, an NFκB signaling pathway, a stress signaling pathway, an apoptosis signaling pathway, an NFAT signaling pathway, a Wnt signaling pathway, a CREB signaling pathway, an AP-1 signaling pathway, a proliferation signaling pathway and an anti-proliferation signaling pathway.

12. A method for conducting a biological readout assay used to screen a bar-coded cDNA library comprising:

(a) sorting the bar-coded cDNA library using a nucleic acid array having a plurality of concave loci;

(b) expressing the bar-coded cDNA library sorted in step (a) using *in vitro* transcription and translation to produce a population of proteins; and

(c) screening the population of proteins produced in step (b) for an activity-of-interest,

so as to conduct the biological readout assay.

13. The method of Claim 12, wherein the activity-of-interest screened in step (c) is selected from the group consisting of a receptor-binding activity, a ligand-binding activity and a growth factor activity.

14. The method of Claim 12, wherein screening is carried out by immobilizing the population of proteins on a solid support and conducting a binding assay with the immobilized population of proteins.

15. The method of Claim 14, wherein the solid support is formed of nitrocellulose or nylon.

16. The method of Claim 12, wherein screening is carried out by placing the population of proteins in contact with readout cells and conducting a biological activity assay.

17. A method for identifying one or more genes-of-interest in a pre-sorted cDNA library comprising:

(a) transfecting the pre-sorted cDNA library into a population of readout cells; and

(b) screening the population of transfected readout cells in a biological readout assay,

to identify one or more genes-of-interest.

18. The method of Claim 17, wherein the pre-sorted cDNA library comprises a bar-coded cDNA library hybridized to a nucleic acid array.

19. The method of Claim 17, wherein transfecting is carried out using chemical transfectants or electroporation.

20. The method of Claim 17, wherein the biological readout assay identifies one or more genes-of-interest in a pathway selected from the group consisting of a mitogenic signaling pathway, a STAT signaling pathway, an NFκB signaling pathway, a stress signaling pathway, an apoptosis signaling pathway, an NFAT signaling pathway, a Wnt signaling pathway, a CREB signaling pathway, an AP-1 signaling pathway, a proliferation signaling pathway and an anti-proliferation signaling pathway.

21. A method of expression cloning one or more genes-of-interest in a cDNA library comprising:

(a) sorting the cDNA library;

(b) transfecting the sorted library into a readout cell line; and

(c) identifying a positive signal from the transfected library in a biological readout assay,

so as to expression clone one or more genes-of-interest in the cDNA library.

22. The method of Claim 21, wherein sorting the cDNA library is carried out using a nucleic acid array.

23. The method of Claim 21, wherein transfecting the sorted library is carried out using chemical transfectants or electroporation.

24. The method of Claim 21, wherein the positive signal is identified by immunocytochemistry.

25. The method of Claim 21, wherein the biological readout assay identifies one or more genes-of-interest in a pathway selected from the group consisting of a mitogenic signaling pathway, a STAT signaling pathway, an NFκB signaling pathway, a stress signaling pathway, an apoptosis signaling pathway, an NFAT signaling pathway, a Wnt signaling pathway, a CREB signaling pathway, an AP-1 signaling pathway, a proliferation signaling pathway and an anti-proliferation signaling pathway.

26. A method of sorting a cDNA library for use in an expression cloning assay comprising:

(a) cloning a population of cDNA inserts into a population of bar-coded vectors;

(b) preparing the population of bar-coded vectors for hybridization to a DNA array by exposing only the bar code region in single-stranded form; and

(c) hybridizing the population of bar-coded vectors to a nucleic acid array to sort the cDNA library.

27. The method of Claim 26, wherein the nucleic acid array is selected from the group consisting of a gene chip and a biological array.

28. The method of Claim 26, wherein preparing the population of bar-coded vectors for hybridization to a DNA array by exposing only the bar code region in single-stranded form in step (b) is carried out by a method comprising the following steps in the order stated:

(a) digesting the population with a restriction endonuclease to linearize the population;

(b) binding a DNA-binding protein to at least two sites on the population; and

(c) digesting the population bound in step (b) to expose the single-stranded bar code region.

29. The method of Claim 28, wherein the DNA-binding protein is selected from the group consisting of a lactose repressor protein, a tetracycline repressor protein, E2F, AP1, SP1 and p53.

30. The method of Claim 28, wherein the restriction endonuclease is selected from the group consisting of NotI, SfiI and EcoRI.

31. The method of Claim 28, wherein digesting the vector population in step (c) is carried out using an enzyme selected from the group consisting of exonuclease III, T4 DNA polymerase, Klenow fragment, T7 DNA polymerase, Vent DNA polymerase and Pfu DNA polymerase.

32. A method for conducting a biological readout assay used to screen a bar-coded cDNA library comprising:

(a) sorting the bar-coded cDNA library using a nucleic acid array having a plurality of concave loci;

(b) contacting the nucleic acid array to a readout cell line in the presence of a solution which facilitates release of the bar-coded cDNA library from the nucleic acid array to carry out *in situ* transfection; and

(c) conducting the biological readout assay.

33. The method of Claim 32, wherein the nucleic acid array is a biological array or a gene chip.

34. The method of Claim 33, wherein the biological array comprises a population of vectors, each vector containing a different bar code complementary to a bar code of the cDNA library to form a population of complementary bar codes, wherein the population of vectors is immobilized on a support.

35. The method of Claim 34, wherein the population of complementary bar codes consists of from 10^2 to 10^8 complementary bar codes.

36. The method of Claim 34, wherein the support is formed of nitrocellulose or nylon.

37. The method of Claim 32, wherein transfecting *in situ* is carried out using a chemical transfectant or electroporation.

38. The method of Claim 32, wherein the readout cell line is NIH 3T3 cells carrying a reporter gene under the control of a response element or promoter.

39. The method of Claim 38, wherein the reporter gene is selected from the group consisting of β-galactosidase, luciferase and chloramphenicol acetyltransferase.

40. The method of Claim 38, wherein the response element or promoter is selected from the group consisting of an NFκB response element, an NFAT response element, a cyclic adenosine monophosphate response element, a STAT-inducible promoter, a LEF-1-inducible promoter and a p53-inducible promoter.

41. The method of Claim 32, wherein expression of the bar-coded cDNA library is tetracycline inducible or estrogen inducible.

42. The method of Claim 32, wherein the biological readout assay is capable of detecting genes in a pathway selected from the group consisting of a mitogenic signaling pathway, a STAT signaling pathway, an NFκB signaling pathway, a stress signaling pathway, an apoptosis signaling pathway, an NFAT signaling pathway, a Wnt signaling pathway, a CREB signaling pathway, an AP-1 signaling pathway, a proliferation signaling pathway and an anti-proliferation signaling pathway.

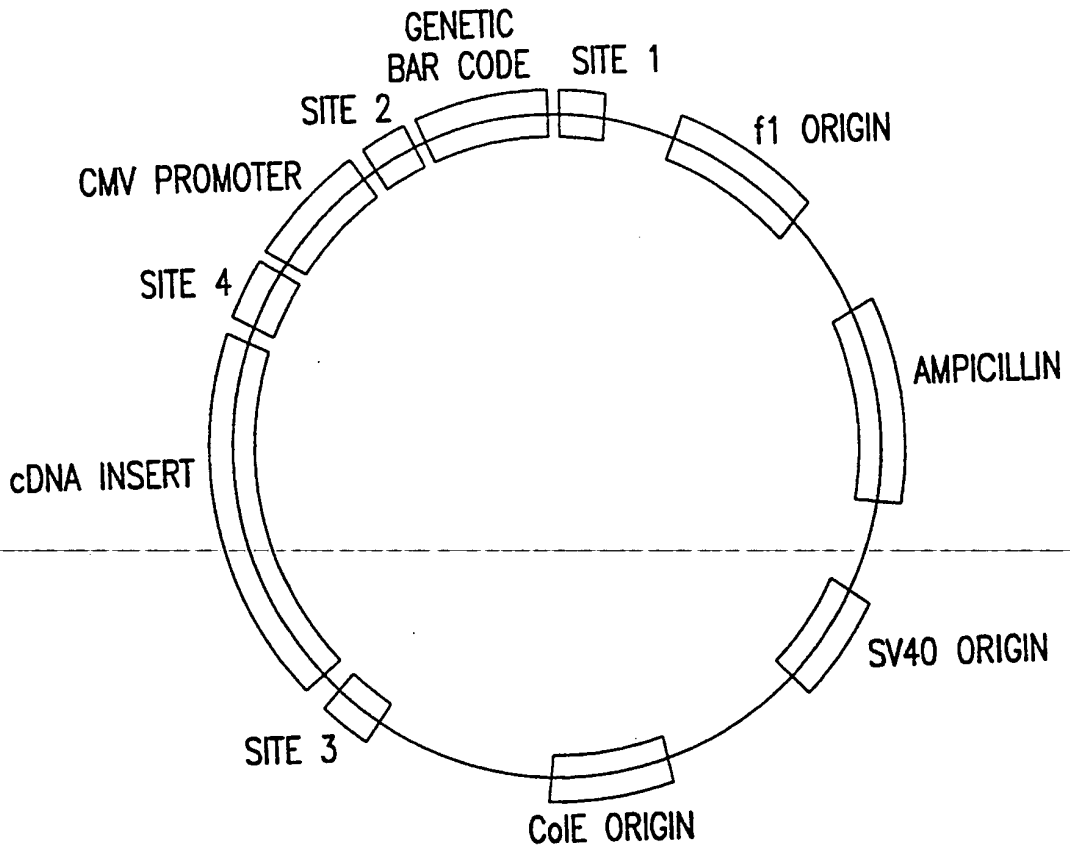

GENETIC
BAR CODE
SITE 1
SITE 2
f1 ORIGIN
CMV PROMOTER
SITE 4
AMPICILLIN
cDNA INSERT
SV40 ORIGIN
SITE 3
ColE ORIGIN

FIG.1

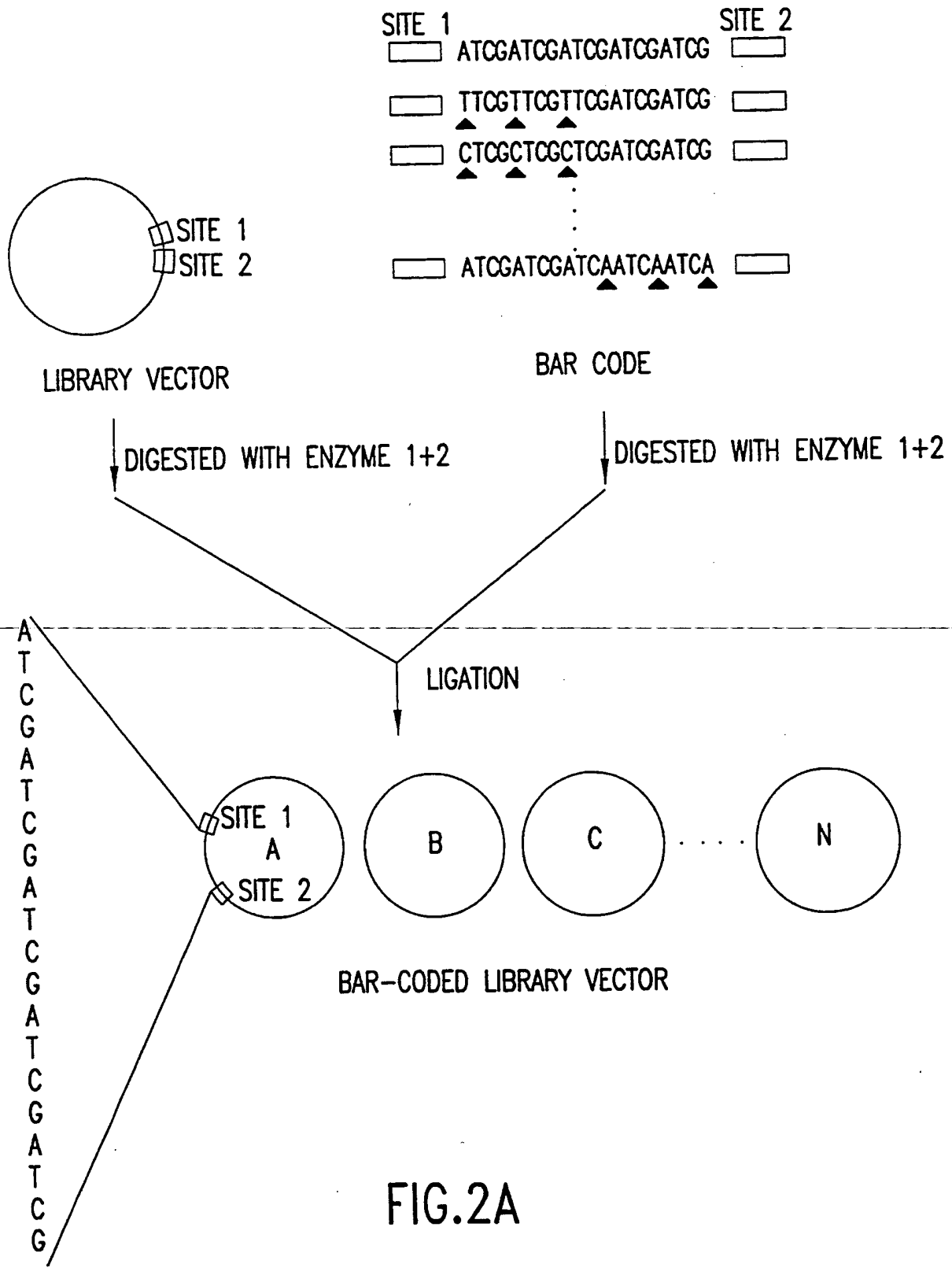

SITE 1
SITE 2
ATCGATCGATCGATCGATCG
TTCGTTCGTTCGATCGATCG
CTCGCTCGCTCGATCGATCG
ATCGATCGATCAATCAATCA
SITE 1
SITE 2
LIBRARY VECTOR
BAR CODE
DIGESTED WITH ENZYME 1+2
DIGESTED WITH ENZYME 1+2
LIGATION
ATCGATCGATCGATCGATCG
SITE 1
A
SITE 2
B
C
N
BAR-CODED LIBRARY VECTOR

FIG.2A

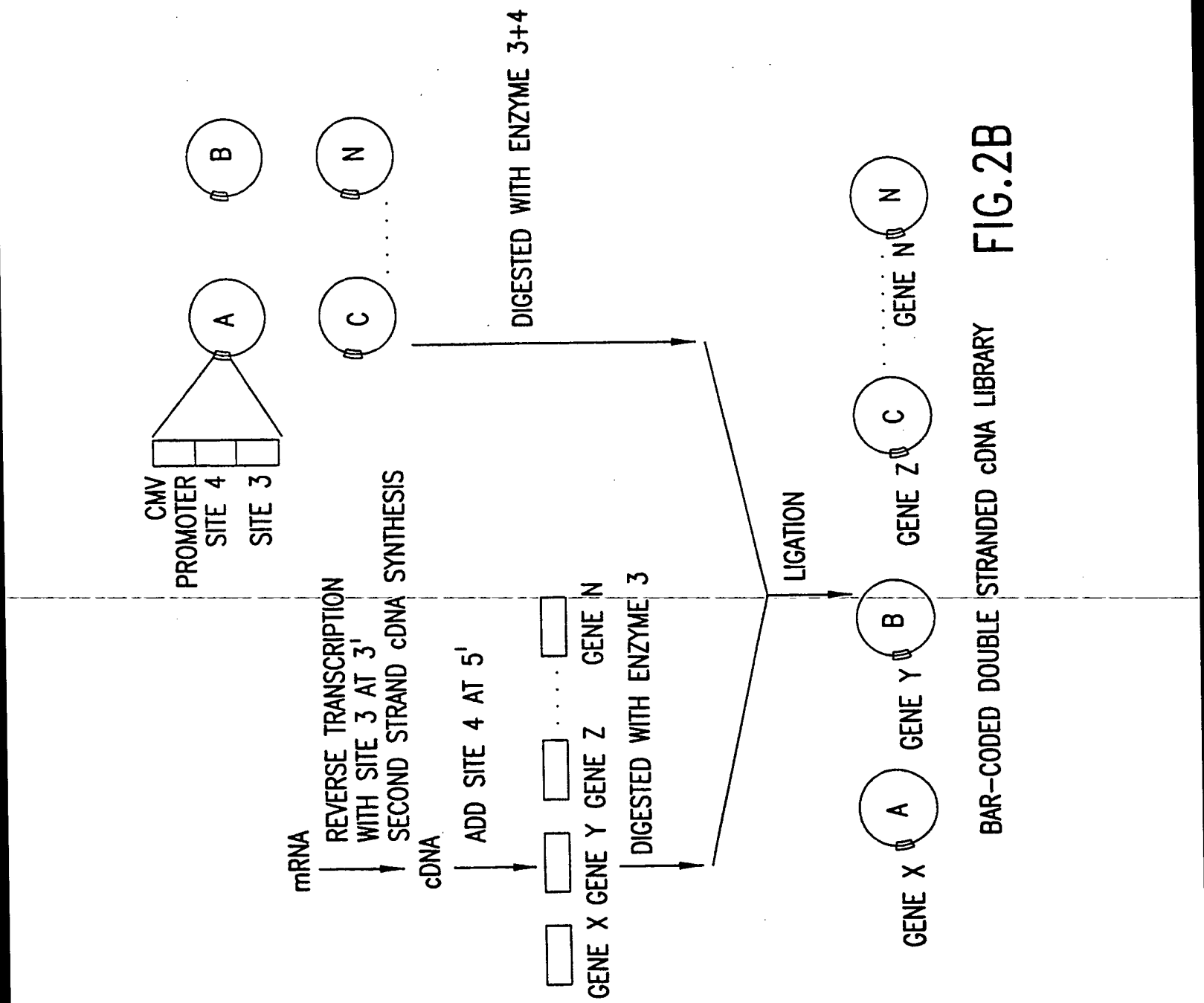

mRNA
REVERSE TRANSCRIPTION WITH SITE 3 AT 3'
SECOND STRAND cDNA SYNTHESIS
cDNA
ADD SITE 4 AT 5'
GENE X GENE Y GENE Z GENE N
DIGESTED WITH ENZYME 3
CMV PROMOTER
SITE 4
SITE 3
A
B
C
N
DIGESTED WITH ENZYME 3+4
LIGATION
GENE X
GENE Y
GENE Z
GENE N
BAR-CODED DOUBLE STRANDED cDNA LIBRARY

FIG.2B

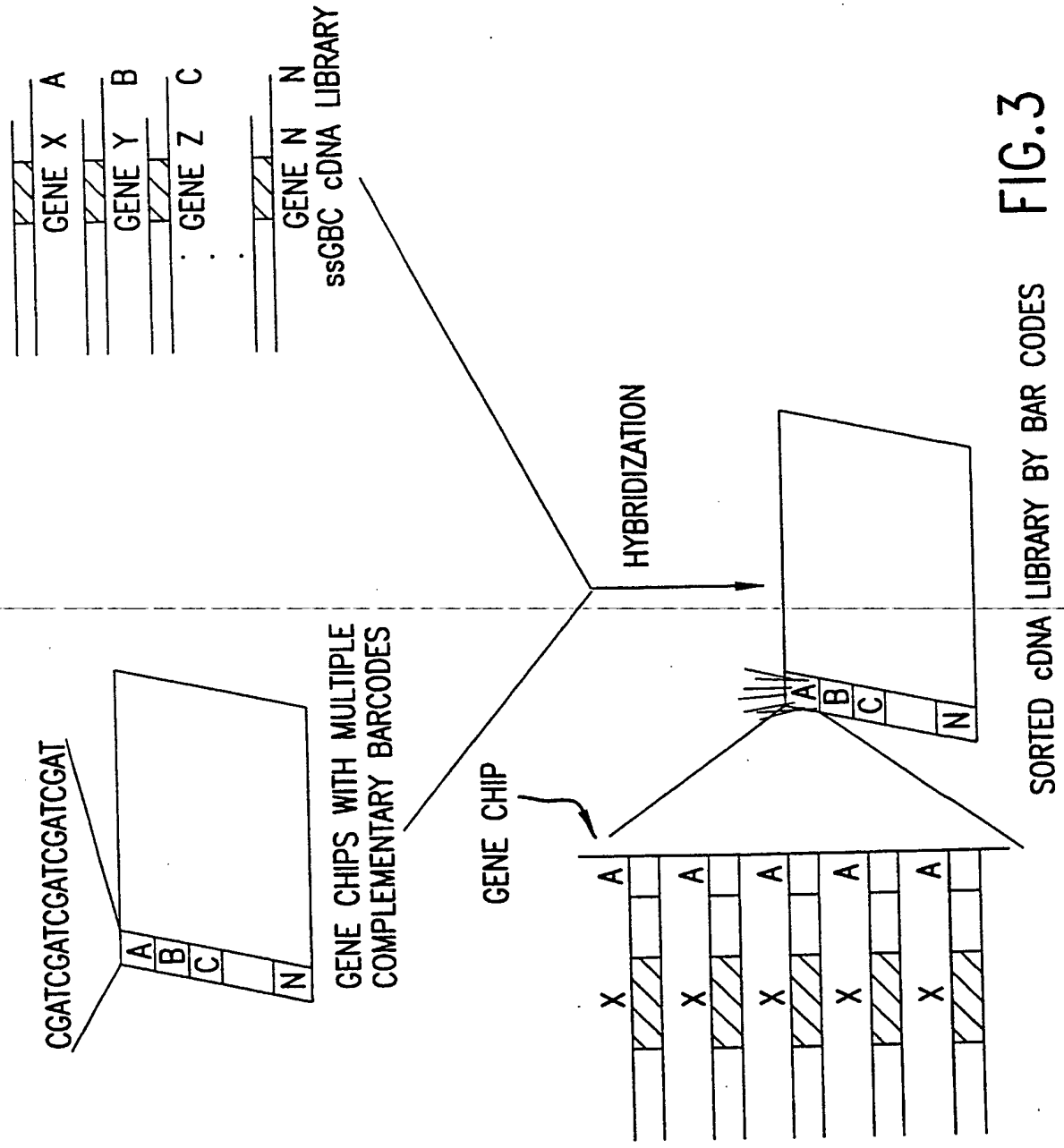

GENE X A
GENE Y B
GENE Z C
GENE N N
ssGBC cDNA LIBRARY
HYBRIDIZATION
CGATCGATCGATCGATCGAT
A
B
C
N
GENE CHIPS WITH MULTIPLE COMPLEMENTARY BARCODES
GENE CHIP
X
A
SORTED cDNA LIBRARY BY BAR CODES

FIG.3

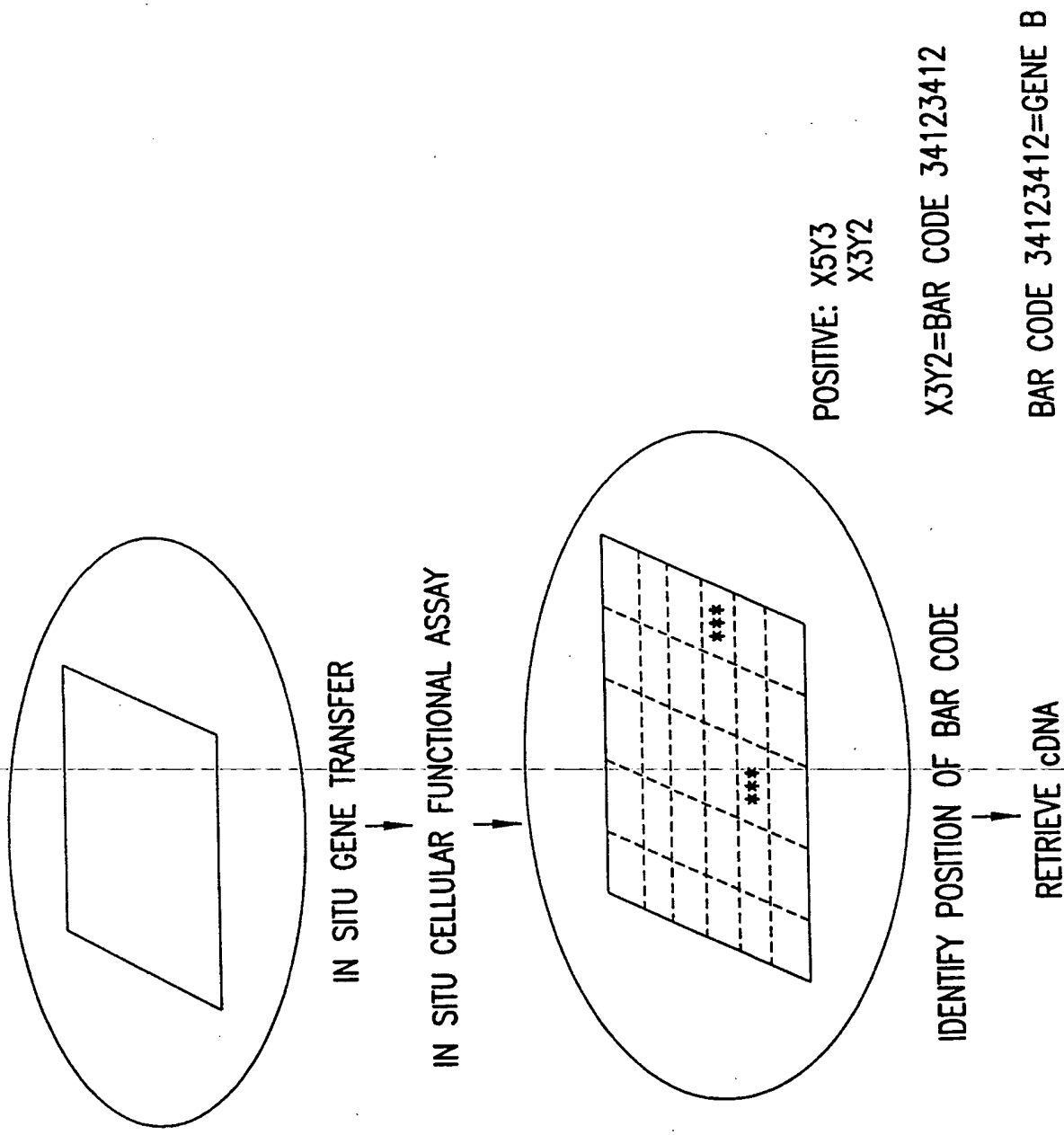
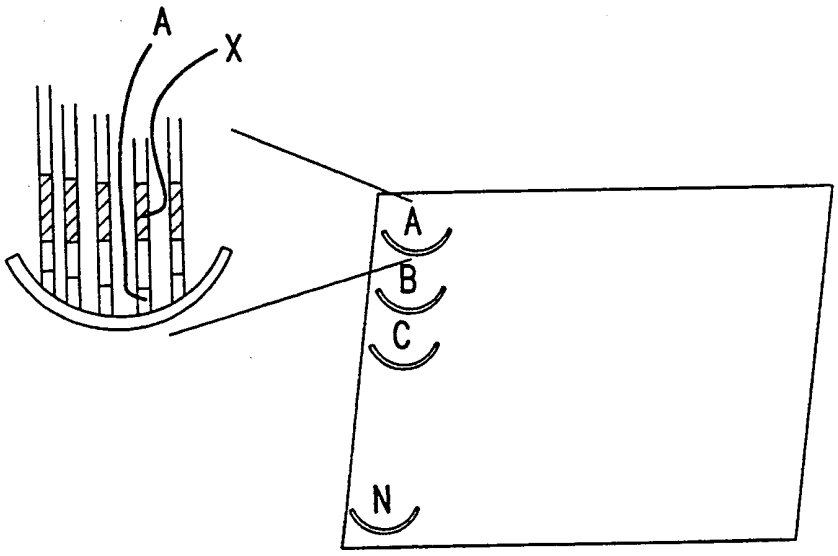

IN SITU GENE TRANSFER
IN SITU CELLULAR FUNCTIONAL ASSAY
IDENTIFY POSITION OF BAR CODE
RETRIEVE cDNA
POSITIVE: X5Y3
X3Y2
X3Y2=BAR CODE 34123412
BAR CODE 34123412=GENE B

FIG.4



GENE CHIP WITH CONCAVE LOCI

X: A cDNA RECOMBINANT
A: A GENETIC BAR CODE REGION

FIG.5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US99/08823

A. CLASSIFICATION OF SUBJECT MATTER
IPC(6) :C12N 15/64, 15/66; C12Q 1/02, 1/68
US CL :435/6, 29, 91.41
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6, 29, 91.41

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Please See Extra Sheet.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 5,445,934 A (FODOR et al) 29 August 1995, see entire document.	1-42
A	US 4,675,285 A (CLARK et al) 23 June 1987, see entire document.	1-42
A	US 5,654,150 A (KING et al) 05 August 1997, see entire document.	1-42
A	US 5,604,097 A (BRENNER) 18 February 1997, see entire document.	1-42
A	EP 0 534 619 A2 (RHODE ISLAND HOSPITAL) 31 March 1993, see entire document.	1-42

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search
13 JULY 1999

Date of mailing of the international search report
30 AUG 1999

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231
Facsimile No. (703) 305-3230

Authorized officer
TERRY A. MCKELVEY
Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)*

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US99/08823

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	EP 0 799 897 A1 (AFFYMETRIX, INC.) 08 October 1997, see entire document.	1-42
A	SHOEMAKER et al, Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. Nature Genetics. December 1996, Volume 14, pages 450-456, see entire document.	1-42

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US99/08823

B. FIELDS SEARCHED
Electronic data bases consulted (Name of data base and where practicable terms used):

APS, DIALOG OneSearch biotech databases
search terms: oligonucleotide?, tags, tagging, assay?, sort?, bar, code?, cellular, cell?, transfect?, transform?, array?, population?, cDNA, librar?, pre-sort?, hybridiz?, expression cloning